XVII AISV CONFERENCE

Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications

Thursday 4th - Friday 5th February 2021

Book of abstracts

XVII AISV Conference

Associazione Italiana Scienze della Voce

Thursday 4th - Friday 5th February 2021

Hosted by University of Zurich

(online)

Organising Committee

Stephan Schmid (chair),

Camilla Bernardasci, Volker Dellwo,

Dalila Dipino, Davide Garassino, Michele Loporcaro,

Stefano Negrinelli, Elisa Pellegrino,

Dieter Studer-Joho

Student Assistant

Seraina Nadig

Scientific Committee

Table of contents

Plenary Lectures

HELEN FRASER *(University of New England, New South Wales, Australia)*

Forensic transcription: Scientific and legal perspectives

Biography. Helen's background is in cognitive phonetics. Since the 1990s, she has undertaken research and case work in forensic transcription, gradually raising concern about injustice arising from the legal handling of indistinct speech recording. She is now Director of the Research Hub for Language in Forensic Evidence, recently established by the University of Melbourne.

Abstract

Covert recordings admitted as evidence in criminal trials are often indistinct. Many jurisdictions allow a transcript, typically produced by police, to help the court understand what is said and who is saying it. Legal procedures include safeguards intended to mitigate the risk of police opinions misleading the trier of fact. However, these safeguards are not always effective. To assist, the scientific community offers acoustic analysis and audio enhancement. While these techniques may be valid in themselves, they can have unintended consequences when they enter the legal process. This presentation gives examples from Australian criminal trials, discusses implications for phonetic science, and outlines solutions being pursued via the new Research Hub for Language in Forensic Evidence.

KIRSTY MCDOUGALL *(University of Cambridge, UK)*

Ear-Catching versus Eye-Catching? Some Developments and Current Challenges in Earwitness Identification Evidence

Biography. Kirsty McDougall is a Lecturer in Phonetics at the University of Cambridge, UK, and a Fellow of Selwyn College, Cambridge. Her research interests range across speaker characteristics, forensic phonetics, theories of speech production and the phonetic realisation of varieties of English. Among other things, her forensic phonetic research has focused on speaker-characterising properties of dynamic features of speech, perceived voice similarity and its implications for voice parade construction, and the development of techniques for analysing individual differences in disfluency behaviour.

Abstract

Earwitness identification evidence may be called on if a perpetrator's voice has been heard at the scene of a crime, but not recorded. If a suspect has been located, a voice parade containing a sample of the suspect's voice alongside a selection of foil voices may be constructed and the witness asked if they are able to recognise the perpetrator's voice in the parade. Earwitness identification of this sort can constitute crucial evidence, yet there remain many unanswered questions about the phonetic and psychological underpinnings of this type of identification and about the optimal way to collect such evidence. In England and Wales, the current procedure for carrying out a voice parade is outlined in a Home Office circular published in 2003. This procedure was developed as an extension of the police procedure for visual identification parades, as informed by the research on earwitness behaviour available at the time. However, developments in psychological research show that while there are some similarities between the processing of faces and voices, considerable differences exist, and further research is needed to determine the best settings of the relevant variables in voice parades (e.g. length of voice samples, number of foil voices, witness instructions, parade type) to minimise earwitness errors. This talk will consider recent findings from the IVIP ('Improving Voice Identification Procedures') Project which is investigating ways the Home Office procedure might be adapted to optimise earwitness performance. The talk will also consider issues relating to the selection of foil voices and the impact of different accents on listeners' perception of voice similarity.

General Session

Do you remain the same speaker over 21 recordings?

Nicolas Audibert, Cécile Fougeron and Estelle Chardenon
Laboratoire de Phonétique et Phonologie
UMR7018 CNRS/Sorbonne Nouvelle, Paris, France
{nicolas.audibert, cecile.fougeron, estelle.chardenon}@sorbonne-nouvelle.fr

Introduction: In forensics, an open question concerns the validity of a comparison between recordings weeks, months, or years apart, and which conditions allow such comparisons. While variation between speakers or speech conditions has been the focus of many phonetic studies, our knowledge on intra-speaker variability across multiple recordings of the same task is surprisingly very limited. Among other factors, age (Biever and Bless, 1989; Jacewicz, 2009), quality of life (Campbell, 2009; Verdonck, 2004), fatigue and emotional state (Scherer et al, 1998; Hollien, 1990), as well as the speech task (Dellwo, 2015) or communicative situation (Scarborough and Zellou, 2013) are known to induce differences in the speech produced by the same individual. More recently, Chardenon (2020) showed that intra-speaker variability on temporal dimensions is larger between distant recording than between successive recordings. This work is part of a larger project on methodological issues in voice comparison, with a specific focus on intra-speaker speech variation across multiple recordings and on the effect of the time lapse between recordings. In the study presented here, we evaluate on a limited set of speakers recorded multiple times over 7 years, whether we can observe the emergence of speaker specific profiles of variation when looking at selected speech dimensions.

Method: Ten speakers were recorded each year three times in a row over a period of seven years, and twice a week over a 1-month period on the same speech material enabling controlled phonetic comparisons between the 29 recording sessions. Recorded material includes read and spontaneous speech as well as other speech-like tasks, in a protocol meant to investigate multiple dimensions of speech and voice. The results presented here are based on 5 female and 3 male speakers, all French native, aged 39 to 58 years old at the date of the first recording, living in the region of Paris and belonging to the same social and professional category. We used 18 to 21 of their recordings (3 successive recordings each year during 6 or 7 years) on the reading of the French version of the tale 'The North wind and the sun'. The text was divided in18 predefined chunks of 15 to 24 phonemes each. All 159 recordings were manually segmented into these 18 chunks, as well as in pauses (with a threshold of 200 milliseconds) and speech. Six features were extracted on each chunk using a Praat script. Information related to the temporal organization of speech is captured over 3 domain-sizes through measures of (i) speech rate (with pauses), (ii) articulation rate (without pause), and (iii) a 'voiced ratio' defined as the total duration of voiced segments over the speech duration. Mean speaking F0 and F0 range over each chunk (in semitones) capture information related to voice and intonation, and the slope of the LTAS captures spectral information related to both laryngeal and supra-laryngeal activities. For these six features, in addition to mean values computed per chunk, the fluctuation of a speech feature is estimated by computing normalized differences (hereafter d(featureX)) between consecutive chunks as $|chunk_i - chunk_{i-1}|/((chunk_i - chunk_{i-1})/2)$. Each recording is thus characterized by 12 features.

Results and discussion: In order to test for the effect of the speaker identity, the recording session, and their interaction, a linear mixed effects model was fitted using the lme4 package (Bates et al., 2015) for each feature converted to z-scores, with the chunk identifier as random intercept. While all 12 features except d(LTAS slope) are found to be speaker dependent, a significant effect of the recording session is found only on mean values per chunk. While differences between consecutive chunks vary by speaker and by recording independently, a recording by speaker interaction is found for mean values by chunk. Interestingly, the descriptors that are more stable across recordings are the ones linked to the fluctuation of the speech dimensions from one chunk to the next over the recording.

In order to further understand individual profiles of variation between recordings, normalized variation levels were estimated by computing the standard deviation over all recordings for each speaker on the 12 features, after conversion to z-scores. As illustrated on Figure 1, variation in temporal dimensions (speech and articulation rate as well as voiced ratio) appears as rather homogenous across speakers with similar variability levels for all of them except articulation rate of M01 and M03. Discrepancies between patterns observed for speech and articulation rate can be attributed to differences in the variability of

pauses number and duration, used by some speakers to compensate variations in articulation rate. On the other hand, speakers are far more different from each other regarding features linked to local variation of F0 (variation of F0 range and fluctuation of F0 mean and range). Although larger differences are found on F0-related features between female speakers, no clear sex-specific patterns are observed.

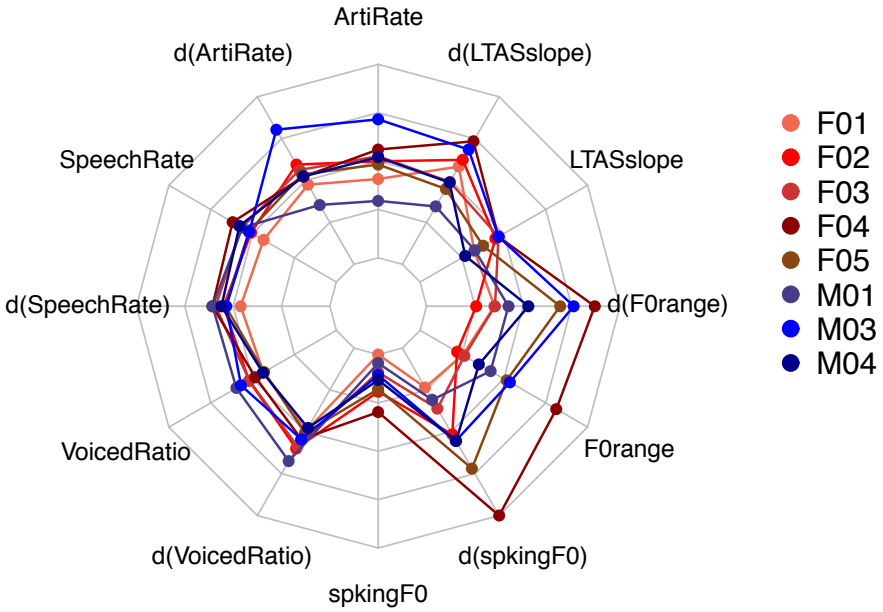


Figure 1: Variation profiles between recording sessions of the 8 speakers, on the 12 speech features expressed as normalized standard deviation for the 6 features measured on the 18 chunks per recording, and their fluctuation between consecutive chunks (chunk to chunk differences are noted as 'd(feature)').

These preliminary results on 8 speakers on the same reading task suggest that measures of fluctuation of speech timing and F0 between consecutive chunks depend more on the speaker than on the repetition. Such measures of local variability, taken at a larger time-scale than the first and second derivatives classically used in automatic classification tasks, may be useful to improve the robustness of speaker identification. They also suggest that patterns of variability between recording sessions are speaker-dependent, particularly on F0 related features. Data recorded for the remaining 2 speakers (1 male, 1 female) are currently being analyzed while the extension of such analysis to spontaneous productions of the same speakers is being investigated. Further analysis will be carried out with more features and a more comprehensive description of the prosodic phrasing of the text in each recording.

References

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).

Biever, D. M., Bless, D. M. (1989). Vibratory characteristics of the vocal folds in young adult and geriatric women. *Journal of Voice*, 3(2), 120-131.

Campbell, W., Bonastre, J-F., Schwartz, R., Driss, M. (2009). Forensic Speaker Recognition. *Signal Processing Magazine*. 26.2: pp-95-103.

Chardenon, E., Fougeron, C., Audibert, N., Gendrot, C. (2020). Dis-moi comment tu varies ton débit, je te dirai qui tu es. *31e Journées d'Études sur la Parole*. Nancy, France, 82-90.

Dellwo, V., Leemann, A., Kolly, M-J. (2015). Rythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *JASA*, 137(3): 1513-1528.

Harrington, J., Palethorpe, S., & Watson, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. *Interspeech 2007*.

Hollien, H., 1990. The acoustics of Crime. (1990). *The New Science of Forensic Phonetics*. Dordrecht: Springer.

Jacewicz, E., Fox, R-A., O'Neill, C., Salmons, J. (2009). Articulation rate across dialect, age and gender. *Lang Var Change*. 1; 21(2): 233–256.

Scarborough, R., & Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *JASA*, 134(5), 3793-3807.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). *Vocal expression of emotion*. Oxford Univ. Press.

Verdonck-de Leeuw, I. M., & Mahieu, H. F. (2004). Vocal aging and the impact on daily life: a longitudinal study. *Journal of Voice*, 18(2), 193–202.

The quest for speaker individuality – a challenge for forensic phonetics

Angelika Braun, University of Trier, Germany

brauna@uni-trier.de

This contribution takes a principled approach towards speaker individuality with the forensic application of phonetics in mind. It is argued from the perspective of a forensic practitioner with extensive casework experience. The question of individuality can be split up into two sub-questions: (a) is "speaking" individual, and (b) can the individuality be detected under forensic circumstances? In examining (b), a further question is how the human listener and the computer deal with detecting speaker individuality. These issues will be addressed consecutively.

Intuitively, there is little doubt that speaking is highly speaker specific and that it is among the elements which lend themselves to be used as a biometric. Indeed, listeners are generally able to recognize familiar speakers even when they whisper or use falsetto voice (Braun / Kraft 2013). The pivot point of this view is the notion of "voiceprint" as an obvious analogy to fingerprint, which was set out by Lawrence Kersta in 1962 (Kersta 1962) and is still widespread among lay listeners today.[1] This fingerprint analogy fails to recognize that – other than fingerprints – speaking is highly variable even within one speaker. Illnesses affecting the speech organs in general and the larynx in particular may render a speaker hard to recognize even by his close friends and relatives. The same is true for extreme emotional states and, e.g., excessive consumption of substances like alcohol or psychodelic drugs.

Speaking forms part of human *behavior* rather than being a mere consequence of anatomical configurations like e.g., the size of the resonance cavities within the vocal tract. Francis Nolan has coined the term *plasticity of the vocal tract* (Nolan 1983, pp. 27-28) in order to describe the within-speaker variation caused by, e.g., pulling the larynx upwards in a stressful situation and thereby reducing the size of the pharyngeal cavity or by protruding one's lips and thereby enlarging the oral cavity.

We can draw the interim conclusion that speaker individuality is still an essentially unresolved issue. There are anatomical limitations within which a speaker can produce "sound"[2]. It is within that range that all speaking behavior takes place. The trouble is, however, that neither the anatomical basis nor the behavioral component is constant. The anatomical/physiological basis changes as soon as the speaker e.g. catches a cold and gets congested; behavioral patterns may change with the situation or a second speaker involved. All of this does not yet include a deliberate change of voice and speech features, i.e. disguise. Consequently, there is a plethora of combinations between speaker anatomy/physiology and behavioral factors including the possibility that between-speaker variability may be smaller than within-speaker variability under certain circumstances (Bolt et al. 1976).

This brings us to the second question, i.e. the robustness of speaker specific features to certain factors which may be present in the forensic setting. In addition to behavioral factors, these include technical issues like telephone transmission or signal reduction by way of coding, such as MP3. Thus the features to rely on in the forensic setting not only have to be speaker specific but also resistant to technical issues like transmission and coding.

[1] The present author was confronted with the fingerprint analogy when recently being interviewed for a popular science TV program, and it proved quite difficult to convince a well-known TV host that the notion of *voiceprint* is invalid.

[2] For reasons of brevity, I take this to include respiratory, phonatory, articulatory and linguistic features alike.

It has been suggested in the past that automatic SR systems may not be particularly sensitive to intraspeaker variability caused by e.g. emotional states because the Mel frequency cepstral coefficients (MFCCs) which many of them rely on are said to represent the vocal tract anatomy. If, however, vocal tract configuration cannot assumed to be a constant, then systems relying on it should be expected not to perform very well if e.g. the speaking situation changes. This actually seems to be the case: Automatic systems are not only extremely susceptible to channel mismatch (Becker 2011), but also to behavioral mismatch and cannot deal with disguise at all (González Hautamäki et al. 2017; 2019). This may lead to totally erratic results and inexplicable errors.

Another issue which is affected by these considerations is the phrasing of conclusions. For about the past 20 years, expressing conclusions in terms of likelihood ratios has been a frequent demand because the traditional probability scales were said to be logically and statistically flawed (e.g. Champod and Meuwly 2000). LRs have a reputation of being more "objective" in that they render one numeric measure based on extensive statistical background data. While this is certainly correct from a strictly statistical point of view, there are practical considerations which raise some questions. One of them is whether LRs are really so unambiguous or whether they would not have to be tailored to the emotional and physiological states of the speaker. This implies that there could be more than one LR per recording, depending on which background data are used. – On a different strand, LRs do not allow for attaching weight to certain findings. For instance, the forensic practitioner will easily identify findings which will effectively rule out identity. These, however, are not reflected as such in the LRs, which may lead to absurd results of, e.g., two speakers from different parts of the country being taken to be one and the same. The trained human listener is in a much better position to compensate for behavioral and technical mismatches as well as consider parameter salience, and s/he will work these into her/his conclusion on a given probability scale. We may after all have to accept the thought that the "objectivity" of LRs can be challenged on some counts and that the much criticized probability scales, which admittedly involve subjective judgement on the part of the expert, do have some forensic merit.

References:

Becker, T. *(*2011): *Automatischer forensischer Stimmenvergleich*. PhD. Diss. Trier.

Bolt, R.H./ Cooper, F.S. / Green, D.M./ Hamlet, S.L./ Hogan, D.L./ McKnight, J.G./ Pickett, J.M. / Tosi, O./ Underwood, B.D. (1979): *On the Theory and Practice of Voice Identification*. Washington, D.C.: National Academy of Sciences.

Braun, A. / Kraft, L. (2013): "Die Erkennbarkeit vertrauter Stimmen bei Verstellung", In: Mehnert, D. / Kordon, U. / Wolff, M. (Hg.): *Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag.* Dresden: TUDpress, pp. 226-233.

Champod, C. / Meuwly, D. (2000): "The inference of identity in forensic speaker recognition", *Speech Communication* 31, 193-203.

González Hautamäki, R., Sahidullah, M., Hautamäki, V., and Kinnunen, T. (2017). "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication* 95, 1–15.

González Hautamäki, R. Hautamäki, V. and Tomi Kinnunen (2019): "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise", *The Journal of the Acoustical Society of America* 146, 693-704.

Jessen, Michael (2008): "Forensic Phonetics", *Language and Linguistics Compass* 2, 1-41.

Kersta, Lawrence G. (1962): "Voiceprint Identification", *Nature* 196: 1253-7.

Nolan, Francis (1983): *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Silvia Calamai, Maria Francesca Stamuli, Alessandro Casellato

Un percorso condiviso per la redazione di un *Vademecum* sulla conservazione, la descrizione, l'uso e il riuso delle fonti orali

Le fonti orali non sono oggetti disciplinari ma rappresentano un metodo di lavoro. Le idiosincrasie e i confini che delimitano ciascun ambito di studio rendono difficile produrre documenti-guida che possano facilitare il lavoro di chi si trova a costruire e a gestire un 'archivio orale', ovvero un archivio costituito prevalentemente o esclusivamente da documentazione audio o audiovideo.

Ad Arezzo, nel febbraio 2019, si è tenuto il XV Convegno dell'Associazione Italiana di Scienze della Voce (AISV), la cui sezione tematica era dedicata proprio alle fonti orali. Nella tavola rotonda finale esponenti della ricerca e delle associazioni (AISV e Associazione Italiana di Storia Orale in primis) erano affiancati da direttori/esponenti degli Istituti centrali dello stato deputati alla conservazione e alla descrizione di fonti orali, insieme a funzionari delle Soprintendenze. Lì si sono tenute le 'prove generali' per individuare un lessico comune: se i linguisti parlano di *corpora*, gli oralisti utilizzano preferibilmente l'etichetta 'fonti orali', mentre gli archivisti usano concetti come 'documento' e 'archivio'; nell'ambito dell'infrastruttura europea di CLARIN, invece, si usa più spesso l'etichetta di *collection*, legata agli ambiti disciplinari della bibliografia, della biblioteconomia e del documentalismo.

Sotto l'impulso delle Associazioni e Istituzioni che lavorano con e/o producono 'fonti orali' (per usare l'espressione degli oralisti), si è dunque creato un gruppo di lavoro finalizzato a mettere in rete associazioni, soggetti produttori di servizi e Istituzioni che, in Italia, sono deputate alla tutela e alla descrizione delle fonti orali. In diversi incontri – succedutisi lungo tutto il 2019 e il 2020 – per la prima volta soggetti differenti e diversificati si sono trovati a dialogare insieme: Associazione Italiana di Storia Orale, Associazione Italiana di Scienze della Voce, Soprintendenza Archivistica e Bibliografica della Toscana, Istituto Piemontese per la Storia della Resistenza e della Società contemporanea 'Giorgio Agosti', Istituto Nazionale Ferruccio Parri, Istituto per la Storia dell'Età contemporanea, Istituto Centrale per gli Archivi, Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, Istituto centrale per i beni sonori ed audiovisivi, Istituto Centrale per il Catalogo e la Documentazione, Centro di Sonologia computazionale del Dipartimento di Ingegneria dell'Informazione dell'Università di Padova, CLARIN-IT, il nodo italiano di CLARIN. Il *Vademecum* risultante è pensato per il ricercatore che lavora sulle fonti orali, anche occasionale e indipendente, per gli Istituti che si trovano a gestire materiale eterogeneo e non abbastanza codificato, per le Associazioni disciplinari i cui membri fanno uso di fonti orali. Il gruppo di lavoro, organizzato e coordinato da Silvia Calamai, Alessandro Casellato, presidente dell'Associazione italiana di storia orale, e Maria Francesca Stamuli, funzionario addetto alla tutela degli archivi con documentazione audio e audiovideo della Soprintendenza archivistica e bibliografica della Toscana, ha prodotto *Linee guida* suddivise in tre differenti sezioni, dedicate, rispettivamente, a

- Conservazione
- Descrizione
- Uso e riuso

Uno dei problemi più critici associati alle 'fonti orali' è la discontinuità e la dispersione non solo della documentazione ma anche delle attività e dei saperi ad essa connessi. Il primo obiettivo del gruppo, quindi, è stato quello di far entrare in sinergia attività e saperi riferiti al trattamento della documentazione sonora e audiovisiva, centrale nella formazione delle 'fonti orali'. Le questioni aperte individuate dai soggetti promotori e dai partecipanti al gruppo fanno evidentemente capo agli aspetti intrinseci di fragilità ampiamente riscontrati ed evidenziati nella letteratura scientifica che a vario titolo e da prospettive metodologiche diverse, ha preso in esame le 'fonti orali':

- FRAGILITÀ MATERIALE: alle difficoltà conservative mediamente superiori rispetto a quelle che afferiscono al 'medium' tradizionale delle fonti documentarie 'classiche' (la carta) si aggiunge una condizione 'intermedia', 'non definitiva' del documento sonoro / audiovisivo: solo recentemente, in molte discipline, si è ha ribaltato il rapporto tra cosa dovesse essere 'mantenuto' (anche per la conservazione futura: usualmente le trascrizioni e non il sonoro) e cosa invece 'scartato' (un tempo, appunto, il sonoro, perché 'meno autorevole').
- FRAGILITÀ EPISTEMOLOGICA: le etichette 'fonte orale', 'documento' e 'archivio' 'sonoro e/o audiovisivo', 'collezione di audiovisivi' sono polisemiche, di livello epistemologico diverso, e dovrebbero essere oggetto di un comune sforzo di definizione per individuare correttamente i loro usi e i loro confini, financo negli aspetti di struttura 'diplomatica' del documento (qual è l'unità informativo-documentaria? coincide o meno con l'unità fisica? in che modo l'unità documentaria e quella fisica possono essere tenute insieme?)
- FRAGILITÀ DESCRITTIVA: le difficoltà descrittive delle 'fonti orali' sono dovute, oltre che alla loro difficile individuazione, anche al fatto che siano trasversali a diverse discipline, caratterizzate da diverse tradizioni di trattamento descrittivo dell'oggetto e diversamente radicate nelle due grandi scienze di descrizione documentale: la biblioteconomia e l'archivistica. Solo all'interno dell'istituzione statale deputata alla tutela dei beni culturali, il Ministero per i beni e le attività culturali, esistono almeno tre tipi di 'schede' di descrizione del documento audiovideo, rispettivamente afferenti al mondo della biblioteconomia, dell'archivistica e della catalogazione demoetnoantropologica.
- FRAGILITÀ GIURIDICO-AMMINISTRATIVA: i dati contenuti nelle fonti orali sono di natura diversa (dati biometrici, dati personali e, non di rado, sensibilissimi; creazioni dell'ingegno esse stesse; metadati); pongono quesiti sulle modalità di escussione dei dati, sull'uso (soprattutto in termini di accessibilità e visibilità) e sul riuso, e chiamano in causa la complessa normativa riferita sia alla tutela della riservatezza e del diritto d'autore.

L'intervento mira dunque a narrare la storia di questa complessa operazione di orientamento e raccordo condotta in Italia: i punti di partenza, le tappe intermedie e i risultati finali.
Tutta la documentazione integrale, dopo la presentazione pubblica del 27 ottobre scorso, è accessibile dai siti di tutti gli enti coinvolti, compreso quello dell'Associazione Italiana di Scienze della Voce. Gli estensori del documento hanno stabilito di 'aprire' a una revisione pubblica il lavoro fin qui condotto: è dunque possibile, attraverso il form reperibile all'indirizzo https://forms.gle/U48rUuk2zDrrig5R7, proporre osservazioni ed emendamenti (entro la scadenza del 15.01.2021).

The Current Situation of the Application of Evidence of Forensic Phonetics in Courts of China

Honglin Cao, Xiaolin Zhang
Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China.
caohonglin@cupl.edu.cn | xiaolin__zhang@163.com

Introduction

Since the *Provision on the Online Issuance of Judgment Documents by People's Courts*[1] passed by the Supreme People's Court of China on Nov. 2013, more than 100 million judgment documents (JD) have been published on the website of *China Judgement Online*[2]. The online JDs provide an irreplaceable resource for big data analysis of numerous justice studies, including research on forensic phonetics (FP). For example, Bao et al. (2016) analyzed 14 JDs related to FP (495 JDs in total for 22 forensic disciplines) which cited the expert opinions given by the forensic practitioners in Institute of Forensic Science, Ministry of Public Security. Cao & Ding (2018) investigated 290 FP-related JDs in courts of six major cities in China. However, the whole application of FP evidence in all nationwide courts is still unknown.

Empirical Survey

In the present survey, 244 effective JDs were downloaded from the website of *itslaw*[3]. The target JDs in FP were selected in the light of these criteria: the regions of courts were restricted to 31 provinces in Mainland China; time was restricted to 2017; key words were restricted to several combinations of the commonest terms in FP, e.g. "recording" + "identification"; the actual-relative JDs in FP were narrowed down through further searching and reading; if more than one JDs were related to a specific report, only one JD was selected as the target JD. Dozens of variables were extracted and analyzed from the target JDs.

Results

Our survey revealed the following results:

(1) Forensic institutes and caseload.
At least 62 forensic institutes/ laboratories in 22 provinces provided services of FP for courts at all levels in 27 provinces. The governmental institutes (22), private institutes (21), and institutes affiliated with universities (19) are roughly equal in number. The top three provinces in terms of number of institutes are Guangdong (9), Beijing (6) and Jiangsu/Liaoning (tied for 5). The number of institutes and the number of cases is positively correlated with the economic development level of different provinces. Among all institutes, the Institute of Forensic Science (Ministry of Public Security, China) provided FP services for at least 11 provinces.

(2) Cause of action.
The criminal, civil and administrative cases accounted for 50.7%, 48.4% and 0.9%, respectively. The criminal cases are predominated by forensic voice comparison (FVC); while the civil cases are predominated by audio authentication (AA). The most criminal cases involve drug crimes, followed by fraud crimes; while the most civil cases involve contract disputes, followed by labor disputes.

(3) Recording methods and devices.
The voice recording methods in the current casework are still predominated by call recording (52.9%), followed by face-to-face direct recording (30.0%) and WeChat voice messages (12.3%). Recording devices are mainly cell phones (67.9%) and digital voice recorder (15.3%).

(4) Forensic tasks.

[1] http://www.chinacourt.org/law/detail/2013/11/id/147242.shtml
[2] http://wenshu.court.gov.cn/
[3] https://www.itslaw.com/

FSC (59.3%) and AA (31.1%) are the first two commonest tasks. Although some tasks are still controversial, like audio originality examination, recording time detection and original recorder examination, they do exist in the court practice, accounting for about 7% of the entire cases. (see figure 1)

(5) Form of expert opinions.
Among the 5-level verbal probability scale framework for FVC, "identification" accounted for 83.1%, followed by "possible identification" (9.2%). For AA, the main form of expert opinion is "the questioned recording is not found to be edited" (72.6%). For audio originality examination, the most common form of expert opinion is "there is no indication that the questioned recording is not the original recording" (50%).

(6) Speaker gender and age.
The ratio of male speakers to female speakers in FVC cases is about 4 to 1. The average (standard deviation) age of male and female speakers is 39.6 (10.4) and 42.2 (12.2) years old, respectively.

(7) Expert witness in court.
The rate of expert witness testifying in court is only about 2% (5/244).

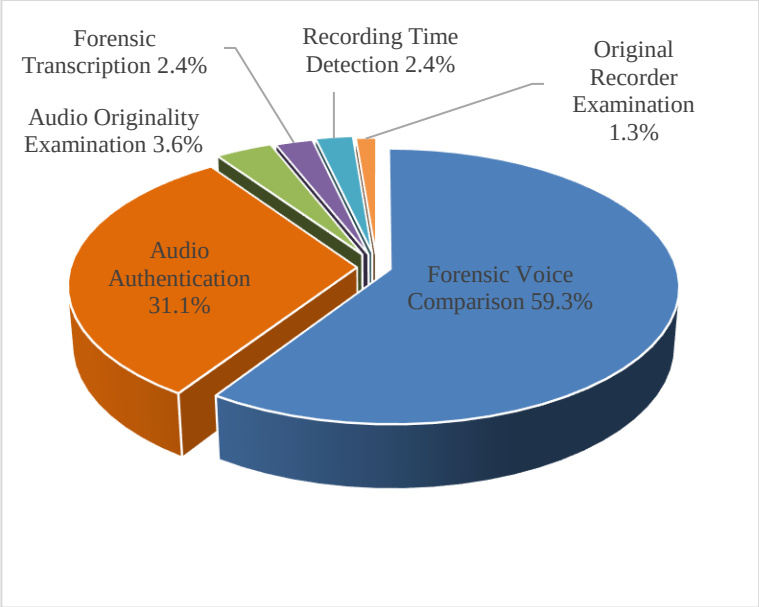


Figure 1. A pie chart with percentage for the six specific tasks.

References

Bao, L., Y. Ge, Y. Jin, B. Cui and K. Ai (2016). Analysis of Expert Testimonial Opinions Quoted in Judgment Documents (in Chinese). *Forensic Science and Technology*., 41(3): 183-188.

Cao, H. and T. Ding (2018). An empirical study on the application of evidence of forensic phonetics in courts of Beijing, Shanghai, Guangzhou, Shenzhen, Tianjin and Chongqing in China (in Chinese). *Evidence Science*., 26(5): 622-638.

The acquisition of /s/ - /z/ in a phonemic vs neutralised context: comparing French$_{L1}$, Italian$_{L1}$ and Spanish$_{L1}$ learners of L2 English

Leonardo Contreas Roa, Paolo Mairano, Caroline Bouzon, Marc Capliez
Université de Lille - UMR 8163 STL

Introduction. English has a high functional load voice contrast between /s/ and /z/, which is active word-initially (*sing* /s/ - *zing* /z/), word-medially (*fussy* /s/ - *fuzzy* /z/) and word-finally (*rice* /s/ - *rise* /z/). However, this contrast is neutralised in the pronunciation of morphemic -s (for plural, 3rd person, genitive, and clitic forms of *has* and *is*). In this specific context, it is subject to a progressive voice assimilation rule (/s/ in *pets* due to /t/ being voiceless, but /z/ in *beds* due to /d/ being voiced) (cf. [1]). In this study we investigate the acquisition of /s/ - /z/ in L2 English by comparing contexts in which these sounds have a phonemic value vs contexts in which they are determined by a voice assimilation rule. We observe English$_{L2}$ productions by French$_{L1}$, Northern Italian$_{L1}$ and Southern American Spanish$_{L1}$ learners, on the assumption that the three groups will show different patterns depending on the status of [s] and [z] in their L1s. These sounds are phonemic in French (*hausse* /os/ - *ose* /oz/), and allophones in varieties of Northern Italian, where [z] appears before voiced Cs and between non-C segments, and [s] in front of voiceless consonants (cf. [2]). Spanish only has /s/ (although partial or total voicing can occur in syllable coda due to non-obligatory voice assimilation with the following C in casual speech, cf. [3]). So, based on SLM predictions (cf. [4]), we expect that (i) French$_{L1}$ learners will show categorically distinct realizations for /s/ and /z/; (ii) Northern Italian$_{L1}$ learners will be able to produce the voice opposition, though potentially to a lesser degree since these sounds are allophones in their L1; (iii) Spanish$_{L1}$ learners may not be able to produce any difference for /s/ and /z/; (iv) voice assimilation for morphemic -s will be difficult for all learners because word-final /z/ is universally more marked (cf. [5]), and all L1s have regressive (rather than progressive) voice assimilation rules.

Data. We analysed productions by 40 instructed learners from the IPCE-IPAC corpus of L2 English. Learners were 15 speakers of Metropolitan French (12 F, 3 M, age = 24, SD = 6.59), 15 speakers of Northern Italian (11 F, 4 M; age = 22.5, SD = 2.38), 10 speakers of Spanish (3 F, 7 M; age = 30.2, SD = 6.98) from Peru (n = 5), Chile (n = 3), Colombia (n = 2). For the present study we only considered recordings of the read-aloud task (506 words), which provides perfectly comparable data. The recordings were transcribed orthographically, phonetized and aligned with *WebMAUS*, and manually verified. For each occurrence of /s/ and /z/, the proportion of periodic signal was extracted via a custom Praat script, thereby obtaining a value ranging from 0 (no periodicity detected) to 1 (periodicity detected throughout the whole target segment). Additionally, we also extracted the duration of segments as a secondary cue of voicing, but durational data are not discussed in this abstract due to space constraints. The results were then imported to *R* for visualisation and statistical analysis.

Results. The results for phonemic /s/ and /z/ (Figure 1) reflect the expected pattern: Spanish$_{L1}$ learners tend not to produce any difference in periodicity between /s/ and /z/, whereas French$_{L1}$ and Italian$_{L1}$ participants show distinct realizations for these two sounds. This is confirmed by a linear mixed-effects model predicting the periodicity on the basis of Sound (/s/, /z/), Group (FR, IT, SP), Context (intervocalic, non-intervocalic) and Position (word-media, word-final). Post-hoc tests confirmed that the difference in periodicity between /s/ and /z/ is significant for French$_{L1}$ and Italian$_{L1}$ learners (p < .001), but not for Spanish$_{L1}$ learners (p = .11). Instead, the results for the pronunciation of morphemic -s (figure 2) show that all learner groups tend to reproduce (at least globally) the output of the voice assimilation rule: the target segments are voiceless when following a C$_{[-voice]}$, and partially periodic if following a C$_{[+voiced]}$ or a vowel. While French learners clearly show the highest differentiation among conditions, it is surprising to observe that Spanish$_{L1}$ learners seem to be able to reproduce the assimilation pattern, and even more neatly than Italian$_{L1}$ learners. This is confirmed by a linear mixed-effects model similar to the one above, revealing that differences across conditions are significant for

the French$_{L1}$ group (all *p* values < .001) and the Spanish$_{L1}$ group (all *p* values < .03), but for the Italian group the condition C$_{[-voice]}$ does not significantly differ from the condition C$_{[-voice]}$ (*p* = .27).

Conclusions. SLM predictions were confirmed in phonemic contexts. Instead, we find unexpected results for the voice assimilation rule, whereby Spanish$_{L1}$ learners manage to produce more voiced realisations than Italian$_{L1}$ learners. The existence of a non-obligatory voice assimilation rule in Spanish (incl. morphemic -s, used for plural in Spanish) may promote voice assimilation in syllable-coda in L2 English, although with a change in directionality (regressive to progressive assimilation). Moreover, the behaviour shown by Spanish$_{L1}$ and Italian$_{L1}$ learners in morphemic vs non-morphemic -s may reflect [6]'s findings for L1 English that morphemic and non-morphemic -s are not homophonous. If confirmed on more data, such results may have an impact on mainstream models of L2 phonology acquisition.

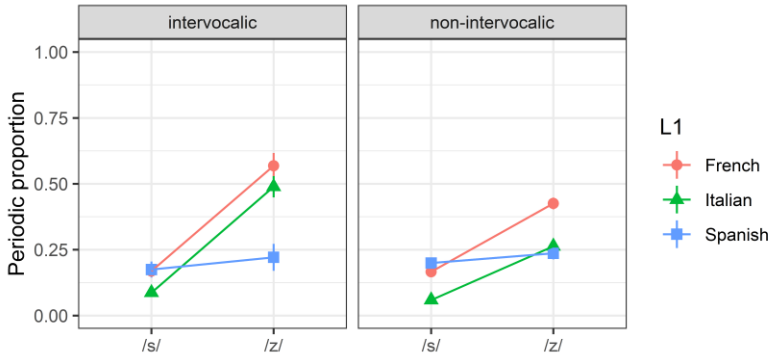


Figure 1. Average periodic proportion for realizations of /s/ and /z/ in phonemic contexts.

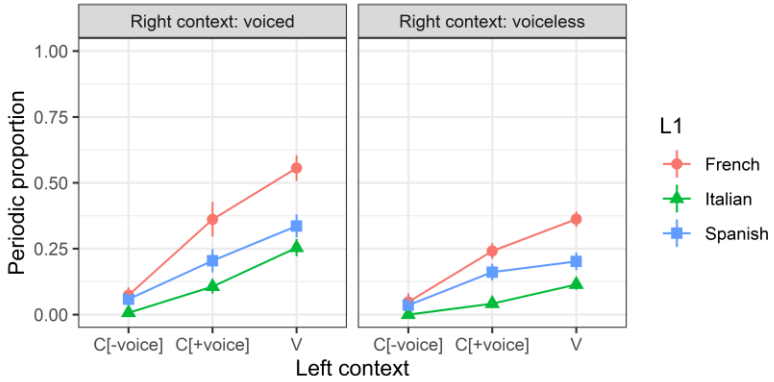


Figure 2. Average periodic proportion for realizations of morphemic -s.

REFERENCES

[1] Cruttenden, A. (2014). *Gimson's pronunciation of English*. Routledge.

[2] Baroni, A. (2014). Element Theory and the Magic of s. *Eugeniusz Cyran–Jolanta Szpyra Kozłowska (szerk.) Crossing Phonetics–Phonology lines, Newcastle upon Tyne, Cambridge Scholars*, 3-30.

[3] Hualde, J. I. (2005). *The sounds of Spanish with audio CD*. Cambridge University Press.

[4] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277.

[5] Eckman, F. R. (2008). Typological markedness and second language phonology. In H. C. Hansen Edwards, M.L Zampini (Eds.) *Phonology and second language acquisition* (pp. 95-115). Amsterdam: John Benjamins.

[6] Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, 53(1), 181–216.

Realizzazione di suoni nativi nel parlato di Italiano L2 da parte di parlanti francofoni: Interazione tra accuratezza e contesto

Sonia d'Apolito, Barbara Gili Fivela
Università del Salento, CRIL-DREAM

In questo studio si vuole osservare, oltre alla relazione tra le caratteristiche del sistema fonetico-fonologico della L1 e della L2 [1,2], l'interazione tra accuratezza nella produzione del parlato e contesto in cui avviene la comunicazione, poiché l'abilità nel parlare in modo accurato, chiaro e con uno stile di eloquio adeguato permette tanto all'apprendente di esprimere il proprio messaggio senza sforzo [3] quanto al percipiente di comprendere il messaggio senza fatica e ambiguità. In base ai dati riportati in letteratura, gli errori fonetici e fonologici hanno un impatto solo del 22% sull'intelligibilità del parlato L2 [3,4], ma il contesto in cui avviene la comunicazione può fortemente influenzare le attese dei percipienti come anche l'accuratezza in produzione da parte degli apprendenti necessaria per esprimere il proprio messaggio [5]. In particolare, in questa indagine studiamo l'accuratezza nella pronuncia di suoni non nativi (geminate) dell'italiano L2 parlato da apprendenti francofoni con due livelli di apprendimento (principiante e avanzato, differenziati in base alle conoscenze pregresse e alla durata della permanenza in Italia, ossia all'entità dell'esposizione alla varietà parlata nell'area di permanenza), al variare della quantità di informazioni presenti nel contesto (contesto più/meno ricco). L'obiettivo è quello di osservare: 1) l'interazione tra le caratteristiche dei sistemi fonetici-fonologici L1 e L2 rispetto alle geminate; e 2) l'interazione tra contesto e accuratezza nella produzione di suoni non nativi. Ci si aspetta che la realizzazione dei suoni non nativi possa variare in base ai seguenti fattori: 1) livello di apprendimento: maggior grado di accuratezza da parte degli apprendenti di livello avanzato; 2) contesto: maggior grado di accuratezza nei contesti con una ridotta quantità di informazioni e, in generale, nei contesti in cui si richiede uno sforzo e un'attenzione maggiore al fine di evitare ambiguità (es. coppie minime).

Otto apprendenti francesi di Italiano L2, quattro principianti e quattro di livello avanzato (in base ai risultati del test Erasmus) e 3 italofoni per controllo hanno partecipato all'esperimento. Quattro coppie minime per ciascun suono target (/l, n, d, r, s, t/) sono state realizzate: 1) in isolamento (es. *ca[s]a, ca[ss]a, entrambe con fricativa sorda nella varietà di riferimento*); 2) in coppia minima (es. *ca[s]a~ca[ss]a*); e all'interno di interazione in contesto che: a) non facilita la disambiguazione (es. Cosa hai detto? -Maria ha detto casa/cassa di nuovo); b) facilita la disambiguazione (es. - *Maria vive vicino al bosco o al mare? – La casa di Maria è vicina al bosco; - Cosa contiene? La cassa contiene le bottiglie di vino*).

I dati acustici sono stati analizzati in PRAAT [6] segmentando i confini della frase, della parola target e ciascun segmento all'interno della sequenza /'C1V1C2V2/ (C2 = C o CC). Si riportano i risultati relativi alla realizzazione delle geminate e alla durata normalizzata (es. durata segmento/durata parola) di V1 e C2 (C o CC). L'analisi statistica è stata effettuata in R, realizzando modelli misti (lme4 [7,8]) nei quali i fattori fissi sono la sequenza, il contesto, il fonema e il livello di apprendimento, e per la variabilità inter-parlante è stato inserito il soggetto come fattore casuale (random intercept; random slope con il contesto). La significatività ($p<0.05$) è stata calcolata utilizzando il Likelihood Ratio test e il test post-hoc di Tukey.

In generale, i risultati mostrano l'influenza della L1 e del livello di apprendimento, poiché i principianti realizzano il maggior numero di degeminazioni. Considerando la durata dei segmenti come indicativa dell'accuratezza della produzione, ci risulta che anche il contesto influenzi la produzione. Dal punto di vista statistico, infatti, per la durata di V1 e C2 risultano significativi i seguenti fattori: a) il tipo di sequenza: le scempie hanno una durata di C2 minore rispetto alle geminate e la vocale precedente, V1, ha una durata maggiore quando è seguita da una scempia

(V1C) piuttosto che da una geminata (V1CC); b) il compito di produzione/contesto: la durata di C2 è minore per entrambi i compiti di interazione (contesto ricco e povero) rispetto alla parola in isolamento e alla coppia minima; effettuando l'analisi solo sui compiti di interazione si riscontra una durata minore in caso di contesto ricco rispetto al contesto povero; per quanto riguarda la durata di V1, entrambi i compiti di interazione differiscono dalla parola in isolamento avendo una durata minore; anche in questo caso, l'analisi solo sulle interazioni mostra una durata minore in caso di contesto ricco; e c) il fonema: per quanto riguarda C2, /l/ e /r/ presentano le durate minori, mentre /s/ e /t/ le durate maggiori; circa V1, la durata è minore quando è seguita da /l, s, t/; d) livello di competenza (principiante, avanzato e nativo): il test è significativo solo per la durata di V1, per la quale mostra una differenza significativa tra i tre i gruppi poiché la durata è minore per i principianti; per quanto riguarda la durata di C2, è stata effettuata un'analisi considerando le geminate e le scempie separatamente. I risultati mostrano che la durata di CC differisce in modo significativo per i tre gruppi di parlanti e il post-hoc indica una durata minore per i principianti, che differiscono in modo significativo dal gruppo degli apprendenti avanzati e dai nativi; nessuna differenza si riscontra invece tra gli apprendenti di livello avanzato e i nativi. Anche nel caso della durata di C2 per le scempie il test è significativo e il post-hoc mostra una differenza tra i due gruppi di apprendenti con i principianti che hanno una durata maggiore.

I risultati mostrano l'influenza della L1 nelle produzioni dei principianti sia rispetto al numero di degeminazioni realizzate che per la durata di V1 e C2. Per quanto riguarda l'interazione contesto-accuratezza, le durate sia della consonante (C2) che della vocale precedente (V1) sono maggiori nei contesti poveri di informazione (come nel caso della parola in isolamento e dell'interazione in contesto povero; le differenze riscontrate a seconda del fonema saranno discusse durante la presentazione) o che richiedono, comunque, una maggiore attenzione per evitare ambiguità (come nel caso delle coppia minime). La discussione dei risultati sarà volta ad approfondire il ruolo dei contesti considerati sull'accuratezza della produzione dei parlanti (con riferimento a [5]), a seconda della differenza tra il sistema L1 e L2 nel caso del fenomeno linguistico considerato (tipo di fonema, geminazione; in linea con quanto suggerito, ad esempio, in [1,2]).

Bibliografia

[1] Flege J., Hillenbrand J. (1984). Limits on pronunciation accuracy in foreign language speech production, JASA, 76, 708-721.

[2] Flege J., Bohn O. & Meador D. (1999). Native Italian speakers' production and perception of English vowels, JASA, 106, 2973-2987.

[3] Zhang, S. (2009). The role of Input, Interaction and Output in the development of oral fluency. In English Language Teaching, Vol. 2 (4), 91-100.

[4] Smith L. E. and Nelson C., L. (2006). World Englishes and issues of intelligibility, The Handbook of World Englishes, Edited by Kachru B., Kachru Y., Nelson, C., Blackwell Publishing Ltd, 428-445.

[5] Lindblom B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle & Marchal (Eds.), Speech Production & Speech Modeling, Dordrecht,403-439.

[6] Boersma P., & Weenink D. (2008). Praat: doing phonetics by computer.

[7] R Core Team, (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL: http:\www.Rproject.org\.

[8] Bates, D., Maechler M., Bolker B., Walker S., (2015). Fitting Linear Mixed-Effects Models using lme4, Journal of Statistical Software, 67(1), 1-48.

Interspeaker variation in anticipatory coarticulation: A whole-formant approach

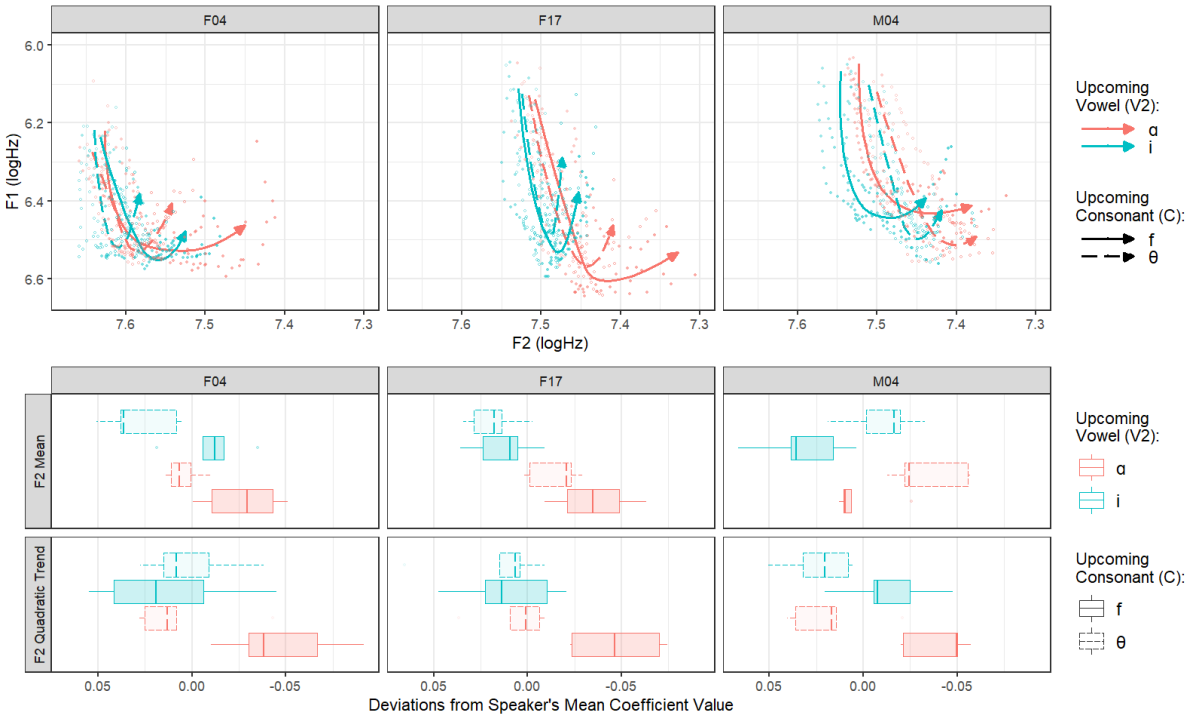
Stefon Flego, Indiana University, Bloomington (sflego@iu.edu)
Jon Forrest, University of Georgia, Athens (jrforrest@uga.edu)

Background: The body of literature studying individual differences in coarticulation identifies variability in a wide range of connected speech processes (Baker et al., 2011; Yu, 2019; Zellou, 2017; Zellou & Pycha, 2018). Most of these studies have made use of static measures of phonetic variation, but the application of time-varying representations of spectral movement to speaker-specific identification (McDougall, 2007; San Segundo & Yang, 2019; Zuo & Mok, 2015) also suggests possible variability between individuals with respect to the dynamics of connected speech processes. Both lines of inquiry suggest that expanding our measurements to include trajectory-based information can shed new light on the fine-grained differences in coarticulatory behavior between speakers. We show that whole-formant representations indeed reveal extensive interspeaker variation in the magnitude and temporal dynamics of anticipatory information in the signal.

Method: Coarticulatory data were collected from 27 speakers of American English who participated in a speech production task. Speakers produced nonce target-context word sequences containing permutations of V_1C#hV_2 within a carrier sentence, in which V_1 = {ɛ}, C = {θ f}, and V_2 = {i ɛ ɑ u}, e.g. *death-heating*, *deaf-hocking*. Timing and prosody were kept consistent across speakers by entraining the syllable rate of the carrier phrase to a metronome. Nonce sequences were chosen instead of existing sequences containing the same target-context vowels to strictly control the prosodic shape of the whole carrier phrase, minimize the number of different lexical items at play, and avoid any lexical frequency effects on coarticulation that may be inherent to existing collocations. The first two formants were measured at the edges of 20 evenly spaced intervals over the course of the target vowel to approximate continuous whole-formant trajectories, to which second-degree orthogonal polynomials were fit. Each estimated orthogonal coefficient provides an independent metric for contour shape (Grabe et al., 2007; McDougall, 2007; Risdal & Kohn, 2014), so their use allows for quantitative analysis of interspeaker variability in formant dynamics.

Findings: Our data not only show a substantial amount of interspeaker variability in basic articulation of [ɛ], but also considerable interspeaker variation in the dynamics of coarticulation with upcoming gestures. To illustrate, average whole-formant trajectories for V_1 = {ɛ} preceding V_2 = {i ɑ} are shown for three of our speakers in the top row of Figure 1, along with distributions of the first and third orthogonal coefficients fit to the F2 contours of these productions in the bottom row. The distribution of the first coefficient reflects gross differences in F2, while the distribution of the third coefficient, the quadratic trend, can be thought of as reflecting different rates of acceleration in F2 over the course of V_1. In addition to differences in overall magnitude of coarticulation observed across individuals, there are robust differences in the phasing of coarticulation, with some speakers showing coarticulatory variation throughout the duration of V_1 (top right panel), while for others these coarticulatory differences only emerge relatively late in V_1's trajectory (top left panel). Furthermore, while variation in F2 attributable to the upcoming vowel is relatively similar across speakers, F2 differences due to the upcoming consonant are largely idiosyncratic (compare the way coefficients cluster with respect to upcoming consonant between M04 and the other two speakers). Finally, the relative contributions of upcoming vowel vs. upcoming consonant to spectral variation differ from speaker to speaker (e.g., the 'F2 Mean' coefficients cluster by consonant for F04 and M04, but by upcoming vowel for F17). These findings support the idea that the temporal dynamics of long-distance coarticulation vary at an individual level. The coefficients of orthogonal polynomial curves fit to whole-formant trajectories provide a means for quantitative comparison of such interspeaker differences.

Figure 1. Top Row: Average formant trajectories for [ɛ], three speakers. Bottom Row: First and third orthogonal coefficients for F2 controlling for speaker means.



References

Baker, A., Archangeli, D., & Mielke, J. (2011). Variability in American English s-retraction suggests a solution to the actuation problem. *Language Variation and Change*, *23*(3), 347–374. https://doi.org/10.1017/S0954394511000135

Grabe, E., Kochanski, G., & Coleman, J. (2007). Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency. *Language and Speech*, *50*(3), 281–310. https://doi.org/10.1177/00238309070500030101

McDougall, K. (2007). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *International Journal of Speech Language and the Law*, *13*(1), 89–126. https://doi.org/10.1558/ijsll.v13i1.89

Risdal, M. L., & Kohn, M. E. (2014). Ethnolectal and generational differences in vowel trajectories: Evidence from African American English and the Southern Vowel System. *University of Penssylvania Working Papers in Linguistics*, *20*(2), 138–148.

San Segundo, E., & Yang, J. (2019). Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation. *Journal of Phonetics*, *75*, 1–26. https://doi.org/10.1016/j.wocn.2019.04.001

Yu, A. C. L. (2019). On the nature of the perception-production link: Individual variability in English sibilant-vowel coarticulation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1), 2. https://doi.org/10.5334/labphon.97

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics*, *61*, 13–29. https://doi.org/10.1016/j.wocn.2016.12.002

Zellou, G., & Pycha, A. (2018). The gradient influence of temporal extent of coarticulation on vowel and speaker perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1), 12. https://doi.org/10.5334/labphon.118

Zuo, D., & Mok, P. P. K. (2015). Formant dynamics of bilingual identical twins. *Journal of Phonetics*, *52*, 1–12. https://doi.org/10.1016/j.wocn.2015.03.003

Prosodia implicita ed esplicita: convergenze e divergenze nella risoluzione di ambiguità sintattiche globali

Salvatore Gianninò[1], Cinzia Avesani[2], Giuliano Bocci[3], Mario Vayra[1,2]
([1]Università di Bologna, [2]ISTC-CNR, [3]Università di Siena)

Introduzione. Nel corso degli anni, le differenze interlinguistiche e intralinguistiche nelle preferenze d'interpretazione delle ambiguità sintattiche globali sono state oggetto di numerosi studi (per una panoramica [9]). In frasi come (1) il secondo sintagma preposizionale (PP2: 'dalla finestra') può costituire un modificatore del sintagma verbale oppure un complemento del sostantivo nel sintagma preposizionale 'dell'evasione' (PP1).

(1)	Verb	Object-NP	PP1	PP2
	Ha visto	i dettagli	dell'evasione	dalla finestra (condizione lunga: 'dalla finestra a ovest')

Nel primo caso si parla di *high attachment* di PP2, nel secondo di *low attachment*. (1) costituisce un caso di ambiguità globale perché le due interpretazioni possibili della frase non sono disambiguate né da informazioni lessicali né da informazioni morfosintattiche. L'unica informazione che si è rivelata cruciale per il *parsing* della frase e la disambiguazione delle due interpretazioni è la struttura prosodica (già [2]), come dimostrato da numerosi studi linguistici e psicolinguistici (per una panoramica [1]).

Janet D. Fodor ha proposto che la prosodia dell'enunciato guidi l'interpretazione della frase non solo quando questa sia pronunciata ad alta voce, ma anche quando sia letta mentalmente. Fodor formalizza la sua proposta nella *Implicit Prosody Hypothesis* (IPH; cfr. [3,4]), con la quale assume che: i) esista una equivalenza tra prosodia esplicita ed implicita; ii) le differenze interlinguistiche rilevate nel *parsing* dei costituenti in frasi del tipo (1) siano riconducibili alla diversa struttura prosodica delle lingue testate; iii) le differenze intralinguistiche siano dovute ad aspetti rilevanti dell'interfaccia prosodia-sintassi, quali, ad esempio, la lunghezza dei costituenti.

Un aspetto tuttora dibattuto della IPH è legato alla natura del *phrasing* prosodico e alla sua relazione col *parsing* sintattico. Possiamo identificare due ordini di problemi.

(1) L'assunto che la prosodia implicita sia uguale alla prosodia esplicita è stato oggetto di dibattito. I dati dei primi esperimenti rilevavano un effetto della lunghezza dei costituenti sul *phrasing* prosodico; quest'ultimo correlava a sua volta con le preferenze di *attachment* [3,9]. Tuttavia, i partecipanti a questi esperimenti leggevano ad alta voce solo dopo una prima lettura silenziosa, quindi quando la struttura sintattica era già formata. Ma se la componente prosodica influisce sulla struttura sintattica, una tale influenza deve aver luogo durante il processo di *parsing*, non quando questo si è già concluso. Nel tentativo di preservare, da un lato, l'equivalenza tra i due tipi di prosodia, dall'altro, il corollario per cui la prosodia esplicita permette di conoscere i tratti della prosodia implicita, si sono condotti diversi esperimenti con lettura ad alta voce all'impronta, senza lettura silenziosa d'anteprima [5,6,7,8,12]. I risultati di questi esperimenti hanno evidenziato un quadro conflittuale: in alcuni casi i confini prosodici erano interpretati come indici sintattici [7,8,12], in altri non lo erano [5,6]. Nel recente studio sull'ebraico di Webman-Shafran e Fodor [12], con un *corpus* molto grande (ca 900 enunciati) ottenuto da letture all'impronta di frasi sintatticamente ambigue, i confini prosodici sono stati interpretati come confini sintattici.

(2) Il secondo problema riguarda *dove* debba essere collocato il confine prosodico nell'enunciato ambiguo per far scattare una o l'altra interpretazione sintattica. In letteratura sono stati proposti due modelli: per l'*Absolute Boundary Hypothesis* (ABH; [11]) sarebbe sufficiente la presenza di un confine prosodico alla sinistra del costituente ambiguo, mentre per la *Relative Boundary Hypothesis* (RBH; cfr. [11]) sarebbe rilevante il rapporto di forza tra questo confine ed eventuali confini prosodici precedenti.

Obiettivi. In questo studio intendiamo verificare la IPH attraverso due studi sperimentali condotti seguendo l'esempio del lavoro sull'ebraico in [12]. Gli obiettivi di questo studio sono: 1) comprendere se la lettura immediata ad alta voce (non preceduta da lettura silenziosa) determini indici prosodici significativi per l'*attachment* e possa, quindi, essere considerata un valido espediente per l'accesso alla prosodia implicita; 2) verificare il ruolo della lunghezza dei costituenti in costrutti globalmente ambigui, per appurare se il *phrasing* prosodico sia influenzato dalla lunghezza dei costituenti e se lo stesso *phrasing* influenzi a sua volta la scelta di *attachment*; 3) verificare il potere predittivo di ABH ed RBH.

Esperimento 1. 20 soggetti italiani, tra i 20 e i 30 anni, hanno letto all'impronta, senza previa lettura mentale, una serie di frasi. Queste includevano 14 coppie di frasi target, sintatticamente ambigue, con struttura sintattica analoga a quella di (1). Ogni coppia comprendeva una frase con PP2 in condizione corta (1 parola prosodica, es. 'dalla finestra') e una frase con PP2 in versione lunga (2-3 parole prosodiche, es. 'dalla finestra a ovest'). Sono stati creati e pseudorandomizzati due blocchi di stimoli controbilanciati, ciascuno con 14 frasi target (7 in condizione lunga e 7 in condizione corta) e 30 frasi filler non ambigue.

Attraverso un software creato allo scopo, i partecipanti visualizzavano su un monitor le frasi, ognuna in un'unica riga, e dovevano leggerle ad alta voce all'impronta. Quindi dovevano scegliere una delle due opzioni interpretative apparse sul monitor. Gli enunciati prodotti sono stati registrati, assieme alle interpretazioni selezionate.

La prima ipotesi è che le frasi *target* in condizione lunga favoriscano un *phrasing* con solo un confine prosodico prima di PP2 e che le frasi *target* in condizione corta favoriscano l'insieme di scansioni con un solo confine prosodico prima di PP1 e senza alcun confine prosodico. Un'ulteriore ipotesi è che il primo di questi *phrasing* determini una prevalenza di *high attachment* e che gli altri due favoriscano invece il *low attachment*. In accordo con [12], all'aumento della lunghezza dei costituenti dovrebbe corrispondere una maggiore frequenza di *high attachment*.

Risultati. Usando la percentuale di errori realizzati nell'interpretazione delle frasi filler come indice di attentività, si sono esclusi i dati ottenuti da 5 partecipanti. I restanti 210 enunciati ambigui sono stati analizzati prosodicamente. Due valutatori hanno identificato percettivamente e in modo indipendente la posizione e l'intensità dei confini prosodici arrivando ad una codifica consensuale di 4 categorie di *phrasing*: presenza di un confine prima di PP2 (*⌋PP2*), di un confine prima di PP1 (*⌋PP1*), di confini in entrambe queste posizioni (*BothBoundary*) o in nessuna di esse (*NoBoundary*). L'analisi statistica con modelli logistici multilivello ha evidenziato un'influenza altamente significativa della lunghezza dei costituenti sul tipo di *phrasing* prosodico prodotto, confermando le nostre predizioni. Le scansioni *⌋PP2* e *⌋PP1*, inoltre, correlano in modo significativo rispettivamente con un *attachment* alto e un *attachment* basso. Tuttavia, la scansione *NoBoundary* non determina una preferenza significativa per un *attachment* basso. L'effetto diretto della lunghezza dei costituenti sintattici sull'*attachment*, infine, non si è rivelato significativo.

Esperimento 2. Un secondo esperimento *web-based* è stato realizzato al fine di rilevare le preferenze di *attachment* con lettura silenziosa. Abbiamo creato diverse coppie di frasi ambigue con una struttura analoga ad (1), manipolando la lunghezza di PP2 come nel primo esperimento. Ricorrendo al giudizio di 3 linguisti esperti rispetto alla plausibilità delle due interpretazioni possibili, si sono selezionate 11 coppie di frasi target. 50 parlanti nativi tra i 20 e i 35 anni hanno preso parte al test. Fatta eccezione per la lettura – stavolta silenziosa – delle frasi, il compito sperimentale rimaneva uguale a quello dell'esperimento 1. Ci aspettiamo un numero significativamente maggiore di *high attachments* per frasi *target* con PP2 lungo e, simmetricamente, un numero maggiore di *low attachments* per frasi *target* con PP2 corto.

Risultati. L'analisi statistica condotta con modelli logistici multilivello ha evidenziato una tendenza conforme alle nostre ipotesi, ma non significativa. Inoltre, indipendentemente dalla lunghezza di PP2, gli *high attachments* prevalgono sui *low attachments*.

Discussione. I risultati dei due esperimenti mostrano che il *phrasing* prosodico risultante da una lettura immediata ad alta voce è influenzato dalla lunghezza dei costituenti (*contra* [5]). Maggiore è la lunghezza del costituente, più probabile è l'inserimento di un confine prosodico al suo confine sinistro. Inoltre, lo stesso *phrasing* prosodico influenza la struttura sintattica di frasi con ambiguità d'*attachment* permanenti (come in [12]), anche se sembrerebbe che ciò si verifichi solo in alcuni casi, in particolare con un *phrasing* di tipo *⌋PP1* ma non con un *phrasing* di tipo *NoBoundary* (cfr. *supra*). Nonostante la IPH risulti scarsamente supportata dai dati ottenuti, la mancanza di effetti della scansione *NoBoundary* sulla struttura sintattica ci fa ipotizzare un quadro più complesso, in cui l'italiano e l'ebraico differiscono in riferimento alla restrizione d'interfaccia *Wrap XP* (Truckenbrodt, 1995; cfr. [10]). Infine, l'analisi acustica – tuttora in corso – degli enunciati rientranti nella categoria *BothBoundary* potrà corroborare i risultati ottenuti tramite giudizi percettivi, permettendo una valutazione quantitativa degli indici acustici correlati ai confini prosodici e la verifica dell'eventuale vantaggio predittivo della RBH sulla ABH.

Riferimenti bibliografici:

1. AVESANI, C., VAYRA, M. (2020), *On the Role of Prosody in Syntactic and Semantic Disambiguation* in E. MAGNI, Y. MARTARI (Eds.), *L'ambiguità nelle e tra le lingue*, special issue of «Quaderni di Semantica», pp. 47-79.
2. CHOMSKY, N., HALLE, M. (1968), *The sound pattern of English*, Harper & Row, New York, NY.
3. FODOR, J. D. (2002a), *Prosodic disambiguation in silent reading*, in M. HIROTANI (Ed.), *Proceedings of NELS 32*, University of Massachusetts, GLSA, Amherst, MA, pp. 112-132.
4. FODOR, J. D. (2002b), *Psycholinguistics cannot escape prosody* in B. BEL, I. MARLIEN, (Eds.), *Proceedings of the Speech Prosody 2002 Conference*, Aix-en-Provence, France.
5. FOLTZ, A., MADAY, K., e ITO, K. (2011) *Order Effects in Production and Comprehension of Prosodic Boundaries*, in FROTA, S., ELORDIETA, G. e PRIETO, P. (a cura di), *Prosodic Categories: Production, Perception and Comprehension. Studies in Natural Language and Linguistic Theory*, Springer, Dordrecht, pp. 39–68.
6. JUN, S.-A. (2010) *The implicit prosody hypothesis and overt prosody in English*, «Language and Cognitive Processes», 25 (7–9), pp. 1201–1233.
7. JUN, S.-A., e KOIKE, C. (2003) *Default Prosody and Relative Clause Attachment in Japanese*, in *Proceedings of the 13th Japanese-Korean Linguistics Conference*, CSLI, Tucson, AZ.
8. JUN, S.A., e KIM, S. (2004) Default phrasing and attachment preference in Korean, in *Proceedings of the 8th International Conference on Spoken Language Processing*, ICSLP 2004, pp. 3009–3012.
9. LOVRIC, N. (2003) *Implicit Prosody in Silent Reading: Relative Clause Attachment in Croatian*, Tesi di dottorato, CUNY.
10. SELKIRK, E. (2000), *The interaction of constraints on prosodic phrasing,* in M. HORNE (Ed.), *Prosody: Theory and Experiment*, Kluwer, Dordrecht, Netherlands, pp. 231–262.
11. SNEDEKER, J., CASSERLY, E. (2010), *Is it all relative? Effects of prosodic boundaries on the comprehension and production of attachment ambiguities*, Language and Cognitive Processes, 25, pp. 7-9, 1234-1264.
12. WEBMAN-SHAFRAN, R., FODOR, J. D. (2015), *Phrase length and prosody in on-line ambiguity resolution*, Journal of Psycholinguistic Research, 45, pp. 447-474.

Do faces speak volumes? A life span perspective on social biases in speech comprehension and evaluation

Adriana Hanulíková (University of Freiburg & FRIAS)

An unresolved issue in social perception concerns the effect of perceived ethnicity on speech processing. Bias-based accounts assume conscious misunderstanding of speech in the case of a talker classification as nonnative (Rubin, 1992; Kang & Rubin, 2009). In contrast, expectation/exemplar-based accounts suggest that correct anticipation of a talker's accent facilitates processing (Babel & Russell, 2015; McGowan, 2015). Driven by theoretical and methodological differences in previous research, this study seeks to establish the extent to which effects of perceived ethnicity on speech processing depend on three sources of variability: experimental method, speech context, and age group. To this end, speech intelligibility and accent ratings from three non-university populations (72 teens, mean age 14.1; 50 younger adults, mean age 36; 50 older adults, mean age 77.6) were examined. Participants were primed with photographs of Asian and White European women and asked to repeat utterances and provide accent ratings for utterances spoken in standard, foreign, and regional accents of German, all embedded in background noise. Repetition accuracy increased when the expected and perceived speech matched, in line with expectation/exemplar-based accounts. This effect varied during the course of the experiment (first vs. second half, see Figure 1) and was most pronounced in the foreign accent and in the group of teens. In contrast, negative effects of ethnicity emerged for accent ratings (see Figure 2) irrespective of the speech context, consistent with a bias-based view. Asian speakers received the most negative accent ratings. The effect was stronger in the group of elderly than in the other groups. Adults showed weak or no effects of ethnicity in either task. The findings show that theoretical contradictions are a likely consequence of methodological choices that tap into distinct aspects of social information processing. Importantly, predictive abilities and strategies vary across the life span, underlining the importance of the inclusion of underrepresented populations in future research.

References

Babel, M., Russell, J. (2015). Expectations and speech intelligibility. *J Acoustical Society of America*, 137(5), 2823–2833.

Kang, O., Rubin, D.L. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *J Lang & Social Psych*, 28, 441–456.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech,* 58(4), 502–521.
Rubin, D.L. (1992). Nonlanguage factors affecting undergraduates' judgements of nonnative English-speaking teaching assistants. *Research in Higher Education*, 33(4), 511–531.

Figure 1: Proportion of correctly repeated words in each speech context and listener group, for the first and second half of the experiment. Black dots represent the overall means and the colored dots show the individual participant means. The violin plots depict probability density. Error bars represent 95% confidence intervals.

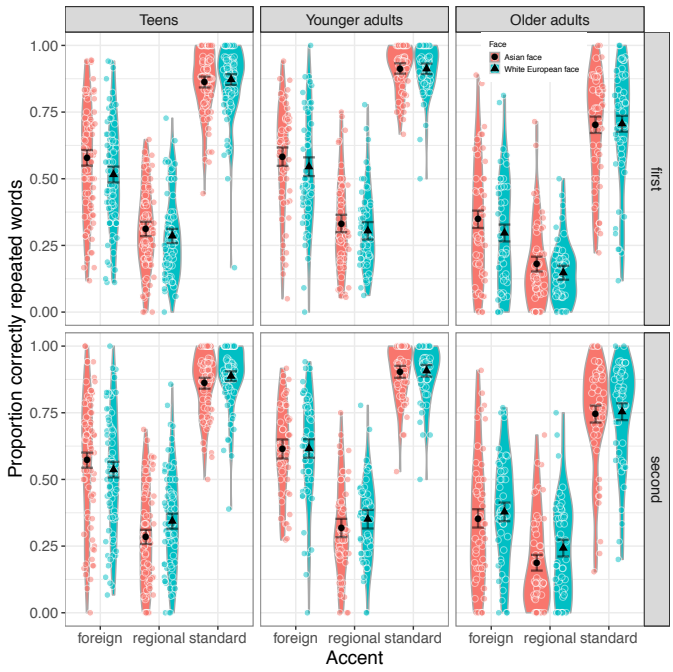
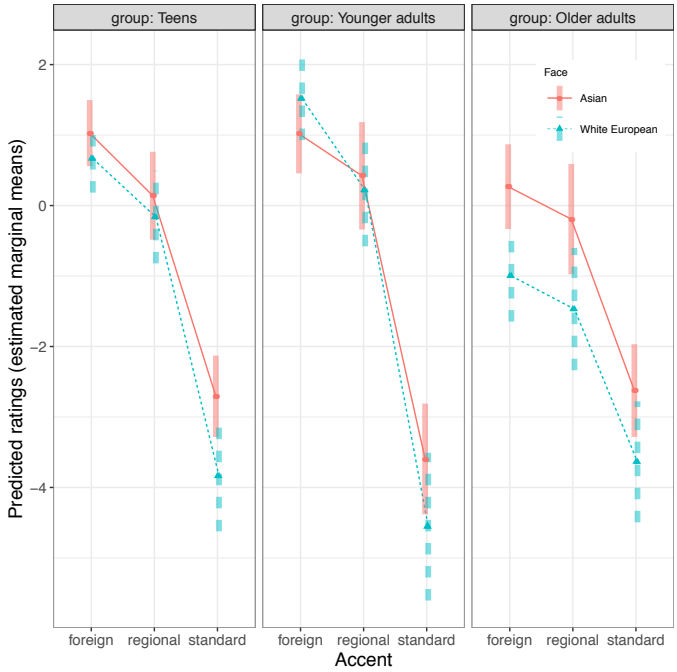


Figure 2: Linear prediction (estimated marginal means) for ratings based on the clmm model. Error bars represent 95% confidence intervals.



Characterizing speech rhythm using spectral coherence between jaw displacement and speech temporal envelope

Lei He
Department of Computational Linguistics, University of Zurich,
lei.he@uzh.ch

Introduction

Lower modulation rates in the temporal envelope (ENV) of the acoustic signal are believed to be the rhythmic backbone in speech, facilitating speech comprehension in terms of neuronal entrainments at δ- and θ-rates (these rates are comparable to the foot- and syllable-rates phonetically) [e.g. 1–3]. The jaw plays the role of a carrier articulator regulating mouth opening in a quasi-cyclical way, which correspond to the low-frequency modulations as a physical consequence. This paper describes a method to examine the joint roles of jaw oscillation and ENV in realizing speech rhythm using spectral coherence. Relative powers in the frequency bands corresponding to the δ-and θ-oscillations in the coherence (respectively notated as **%δ** and **%θ**) were quantified as one possible way of revealing the amount of concomitant foot- and syllable-level rhythmicities borne by both acoustic and articulatory domains. This idea was illustrated using two English corpora (mngu0 and MOCHA-TIMIT) [4, 5] for the proof of concept. %δ and %θ were regressed on utterance duration for an initial analysis. Results showed that the degrees of foot- and syllable-sized rhythmicities are different and are contingent upon the utterance length.

Method

The mngu0 contains one male English speaker producing over 1,000 utterances, amongst which 594 in the duration range of 2–8 sec were chosen for the present study. The 2-sec cutoff allowed at least one cycle of the lowest δ frequency (.5 Hz) to be included; the 8-sec cutoff excluded sentences with medial pauses. The MOCHA-TIMIT (Wrench 1999) contains three English speakers (1f, coded as “fsew0”; 2m, coded as “maps0” and “msak0”) producing the same set of 460 sentences. Altogether 5 sentences shorter than 2 sec were excluded. All utterances were shorter than 6 sec. The EMA data were collected in the meanwhile; of particular interest to this study were the articulatory trajectories of the lower incisor.

Jaw displacements were parameterized as the Euclidean distance of the lower incisor coordinates in the mid-sagittal plane. The ENV were extracted using the full-wave rectification and low-pass filtering. The Jaw-ENV coherence spectra were calculated as the Hermitian inner product of the Fourier coefficients in the FFT of jaw displacement and the FFT of the ENV normalized to the individual power of both FFTs.

The **%δ** and **%θ** were calculated as the percentage of the spectral integral bounded by the δ-band cutoffs ($f_1 = .5$ Hz, $f_2 = 3$ Hz) or θ-band cutoffs ($f_1 = 3$ Hz, $f_2 = 9$ Hz) over the entire spectral integral of the coherence function ($f_{Nyq} = 40$ Hz).

Results

For the mngu0 data, linear regressions between utterance length and %δ and %θ were performed. The utterance duration was right skewed, hence was natural log transformed. %δ increased as utterance duration increased, whereas %θ decreased as utterance length increased (Figure 1).

For the MOCHA-TIMIT data, Random-slope models were fitted by maximum likelihood (response variables: %δ and %θ; random effects: speaker and utterance; fixed effect: utterance length). The significance of the slope estimate and between-speaker variability were tested in particular (Figure 2): in general, a positive slope estimate was found significant between %δ and utterance length, and a negative slope

estimate was found significant between %θ and utterance length. Moreover, individual differences were significant at the same time.

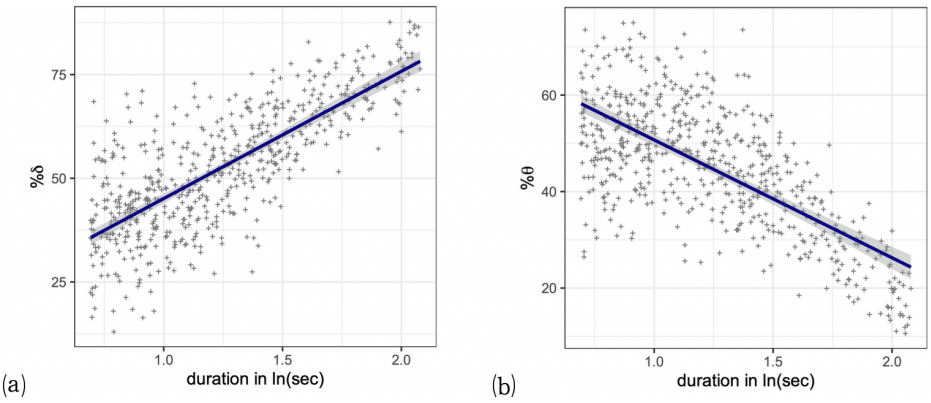


Figure 1: Regression lines and the 99% confidence intervals (shaded areas) superimposed over the scatterplots showing the relationships between %δ and log utterance duration (a), and %θ and log utterance duration (b) in the mngu0 corpus.

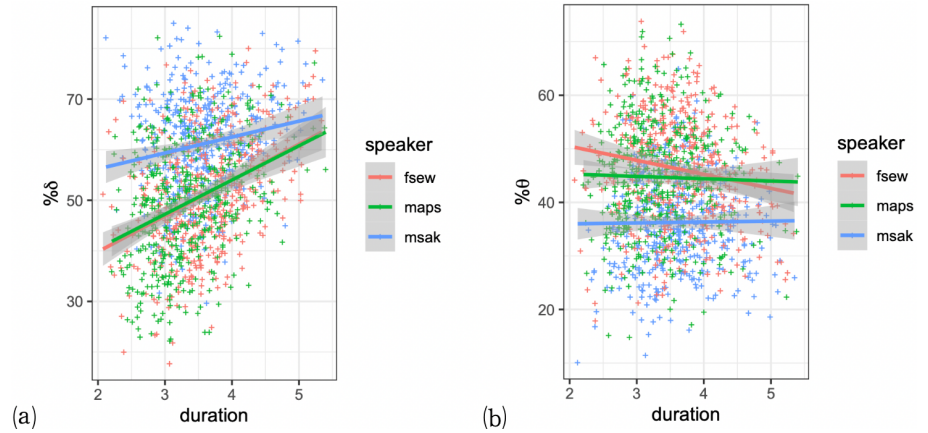


Figure 2: Regression lines and the 99% confidence intervals (shaded areas) superimposed over the scatterplots showing the relationships between %δ and utterance duration (in sec) (a), and %θ and utterance duration (b) for each of the three speakers in the MOCHA-TIMIT corpus.

References

[1] Doelling, Keith B., Luc H. Arnal, Oded Ghitza & David Poeppel. 2014. Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85(2). 761–768.

[2] Ghitza, Oded. 2017. Acoustic-driven delta rhythms as prosodic markers. Language, *Cognition and Neuroscience* 32(5). 545–561.

[3] Poeppel, David & M. Florencia Assaneo. 2020. Speech rhythm and their neural foundations. *Nature Reviews Neuroscience*. 21(6). 322–334.

[4] Richmond, Korin, Phil Hoole & Simon King. 2011. Announcing the electromagnetic articulography (Day 1) subset of the mngu0 articulatory corpus. In *Proceedings of INTERSPEECH 2011*, 1505–1508. Florence, Italy.

[5] Wrench, Alan. 1999. MOCHA MultiCHannel Articulatory database: English (MOCHA-TIMIT). http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

Gender bias in voice recognition: An i-vector-based gender-specific automatic speaker recognition study

Thayabaran Kathiresan[1], Arjun Verma[1], and Volker Dellwo[1]

[1]Phonetics and Speech Sciences, Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

thayabaran.kathiresan@uzh.ch

Abstract

The performance of automatic speaker recognition (ASR) algorithms is continuously increasing. However, acoustic variabilities due to gender differences are challenges, and numerous methods have been proposed to handle them. Among many solutions, using gender-dependent features to train the ASR and testing with known gender trials [1][2] improves the overall recognition accuracy. However, the relative accuracy difference between gender-specific testing still exists. In this paper, we address the fundamental acoustic differences between gender concerning the ASR. We carried out a) an i-vector-based ASR experiment [3] on the TIMIT corpus (130 female and 290 male speakers) and b) the i-vector speaker embedding acoustic analysis. The i-vector extractor was trained on 7323 speakers (1211 speakers from the Voxceleb1 dataset and 6112 from the Voxceleb2 dataset [4]). The system was based on a UBM with 2048 Gaussian mixtures and a gender-independent total variability matrix with 400 total factors. We employed an i-vector length normalization (LN) to the 400-dimensional i-vector. Linear discriminant analysis (LDA) was used to alleviate intra-speaker variability further and reduce the dimension to 200. Finally, PLDA models with 200 latent identity factors were trained. On the trained i-vector system, we carried out the recognition experiment on the male and female speakers separately. The results show that the equal error rate (EER) for male speakers is 3.937% and for female speakers, 5.128%.

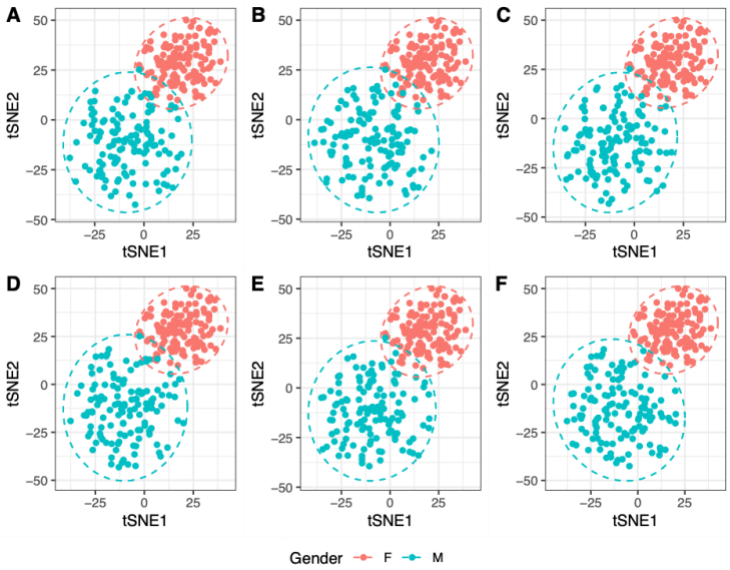


Fig. 1 Distribution of i-vector speaker embeddings (400 dimension) in the t-SNE indexical space (2 dimension). Subfigures A-F indicates the random sampling (6-fold) of male speakers (130 out of 290), and the female speaker count kept constant (130).

We used t-distributed stochastic neighbor embedding (t-SNE), a dimension reduction technique, to reduce the 400-dimensional i-vector to 2-dimension features. To keep the balanced speaker count in both genders, we randomly sampled male speakers to match the female speakers count, and the females' speaker count

was kept constant (Fig. 1). The distributed area of both male and female speaker embedding in the 2-dimensional indexical space was measured, as shown in Fig. 2. The male speakers are relatively sparsely distributed (the mean area is 136.2) than female speakers (the mean area is 49.48).

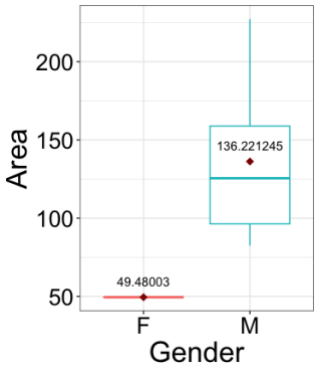


Fig. 2 Area distribution of i-vector speaker embeddings of both male and female speakers in t-SNE indexical space (2 dimension).

We will discuss the results in the context of two plausible explanations: (a) It is possible that voice recognition technology has been developed predominantly based on male voices. Thus, recognition works better for male compared to female voices. Such a gender bias in research knowledge is well known in many scientific disciplines [5], and traditionally research on speech has been predominantly carried out on male voices. (b) It is possible that the acoustics of male voices offer a wider variety of indexical cues to identity and can thus be easier recognized. This would be interesting, particularly from an evolutionary perspective in which voices of different genders could have been attributed different roles in terms of their recognizability [6][7].

References

[1] S. Cumani, O. Glembek, N. Brummer, E. De Villiers, and P. Laface, "Gender independent discriminative speaker recognition in i-vector space," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 1, pp. 4361–4364, 2012, doi: 10.1109/ICASSP.2012.6288885.

[2] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, "Mixture of PLDA models in I-vector space for gender-independent speaker recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 25–28, 2011.

[3] T. Johns, "TIME DELAY DEEP NEURAL NETWORK-BASED UNIVERSAL BACKGROUND MODELS FOR SPEAKER RECOGNITION David Snyder , Daniel Garcia-Romero , Daniel Povey Center for Language and Speech Processing & Human Language Technology Center of Excellence," no. 1232825, pp. 92–97, 2015.

[4] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxceleB2: Deep speaker recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018, doi: 10.21437/Interspeech.2018-1929.

[5] S. Adani and M. Cepanec, "Sex differences in early communication development: Behavioral and neurobiological indicators of more vulnerable communication system development in boys," *Croat. Med. J.*, vol. 60, no. 2, pp. 141–149, 2019, doi: 10.3325/cmj.2019.60.141.

[6] R. Joseph, "The evolution of sex differences in language, sexuality, and visual- spatial skills," *Arch. Sex. Behav.*, vol. 29, no. 1, pp. 35–66, 2000, doi: 10.1023/A:1001834404611.

[7] S. E. Yoho, S. A. Borrie, T. S. Barrett, and D. B. Whittaker, "Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology," *Attention, Perception, Psychophys.*, vol. 81, no. 2, pp. 558–570, 2019, doi: 10.3758/s13414-018-1635-3.

Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition

Katharina Klug, Michael Jessen, Isolde Wagner
Bundeskriminalamt, Germany
{katharina.klug|michael.jessen|isolde.wagner}@bka.bund.de

The research project introduced here intends to expand the application of Forensic Automatic Speaker Recognition (FASR) systems in forensic voice comparison cases by validating the systems using case-specific material. The project is conducted by the German Federal Criminal Police Office (Bundeskriminalamt, BKA), funded by the EU Internal Security Fund (ISF).

Background

The performance of FASR systems improved considerably within the last two decades. Current automatic systems are capable of analyzing even degraded audio recordings forensic audio experts typically face. Therefore, FASR systems have gained wide attention as a component within voice comparison casework among practitioners worldwide (Gold & French 2019).

When case material meets the requirements for FASR application, forensic speech and audio experts at the BKA combine the auditory-acoustic approach with FASR (Wagner 2019). Both approaches follow the principles of similarity and typicality, developed within the Bayesian approach to voice comparison (Rose 2002). The difference between the approaches is that whereas FASR (as well as semiautomatic speaker recognition) generates numerical strength-of-evidence results in the form of likelihood ratios, the auditory-acoustic approach treats similarity and typicality on a more qualitative level (Drygajlo et al. 2015; Jessen 2018). Therefore, if combined, the FASR approach adds quantitative information to the predominantly qualitative information gained from the traditional auditory-acoustic approach.

So far, the case scenarios in which FASR has been applied at the BKA were mainly limited to telephone interceptions (and other forms of natural telephone conversations), with strongest emphasis on speakers of German. These cases generally occur under so-called matching conditions, i.e. both, the recording of the questioned speaker and the one of the suspected speaker (or several of each category), derive from fairly regular telephone conversations (potentially with increased stress and emotion). FASR systems need to be validated on material reflecting case conditions before being used to assess the strength of evidence of a case (Drygajlo et al. 2015). Therefore, the performance of FASR systems in matching telephone interception conditions has been tested at the BKA and in laboratories with similar casework (van der Vloed 2014; Solewicz et al. 2017).

Frequently, however, casework occurs under mismatched conditions, i.e. the technical or behavioural conditions of the questioned speaker's recording(s) differ systematically from those of the suspected speaker's recording(s). Moreover, some casework occurs under matched conditions, but the conditions are not telephone-based or they are telephone-based but have further complications (e.g. high noise level, compression, reverberation). Mismatched conditions and those outside regular telephone conversations are not well represented in validations, which currently limits the scope of cases in which FASR can be used. The few casework-oriented studies on mismatched or unusual conditions that have been conducted recently are based on dedicated recorded speech corpora. Morrison & Enzinger processed high-quality speech recordings in accordance with the mismatched characteristics of a real forensic case (Morrison & Enzinger 2019 for a summary of research that was based on this material). The approach taken in van der Vloed et al. (2020) was to record telephone conversations but under very different technical and environmental conditions that can lead to mismatch or reflect specific limitations.

The research project described in the present paper takes the approach to validate FASR systems using recordings that have occurred in real forensic audio material or as part of related investigative work.

Project

Challenges encountered regarding the compilation of a real forensic audio corpus include:

- highly sensitive data, involving the need to anonymize personal data,
- less-than-complete certainty in determining speaker identification,
- unbalanced number of recordings per speaker,
- unbalanced data for languages and/or conditions of interest.

The table shows the conditions and languages for which audio material has been collected so far.

Table 1. Conditions and languages to be tested with FASR systems (provided sufficient amount of real forensic audio material)

Conditions	Language(s)
Telephone interception	German, Arabic, Turkish, Russian
Video	German, Arabic
Interior surveillance (car/living space)	German, Turkish
Voice message	German

A minimum of 20 male adult speakers per condition and language are collected, providing two to six recordings per speaker. The net duration of speech ranges from 10 to 60 seconds. Additionally, when available, training data sets of independent speakers per condition and language are collected providing only one recording per speaker to create relevant populations. The data preparation includes segmentation into net speech and anonymization of personal information (e.g. names, telephone numbers, addresses).

Using the available data, the performances are being tested of:

- various commercial speaker recognition systems,
- several generations of approaches (GMM/UBM, i-vector, x-vector),
- different methods of score normalization and adaptation.

Current results will be shown during the presentation.

Expanding the FASR application in forensic voice comparison cases will be important for the mismatch issue and for the investigation of matching conditions in non-telephone conditions. The authors hypothesize that the conditions 'video' and 'interior surveillance' will challenge the FASR application most, as speaking styles and recording conditions often vary strikingly within these conditions and typically differ quite strongly from other conditions.

References

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J. & Niemi, T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition.* Frankfurt: Verlag für Polizeiwissenschaft. [http://enfsi.eu/wp-content/uploads/2016/09/ guidelines_fasr_and_fsasr_0.pdf]

Gold, E. & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law 26:* 1-20.

Jessen, M. (2018). Forensic voice comparison. In: J. Visconti (Ed.) *Handbook of communication in the legal sphere.* Berlin: Mouton de Gruyter, pp. 219-255.

Morrison, G. S. & Enzinger, E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01)-Conclusion. *Speech Communication* 112: 37-39.

Rose, P. (200*2). Forensic speaker identification.* London: Taylor & Francis.

Solewicz, Y. A., Jessen, M. & van der Vloed, D. (2017). Null-hypothesis LLR: A proposal for forensic automatic speaker recognition. *Proceedings of Interspeech 2017* (Stockholm, Sweden), pp. 2849-2853.

Van der Vloed, D., Bouten, J. & Van Leeuwen, D. A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. *Proceedings of ODYSSEY 2014* (Joensuu, Finland), pp. 6-13.

Van der Vloed, D., Kelly, F. & Alexander, A. (2020). Exploring the effects of device variability on forensic speaker comparison using VOCALISE and NFI-FRIDA, a forensically realistic database. *Proceedings of ODYSSEY 2020* (Tokyo, Japan), pp. 402-407.

Wagner, I. (2019). Examples of casework in forensic speaker comparison. *Proceedings of the 19th International Congress of Phonetic Sciences* (Melbourne, Australia), pp. 721-725.

Earwitness evidence accuracy revisited: Estimating age, weight, height, education, and geographical origin

Adrian Leemann, Péter Jeszenszky, Carina Steiner, Hannah Hedegard
University of Bern, Switzerland

In forensic casework, there are instances where a victim has heard rather than seen the perpetrator – such as when victims are blindfolded and robbed, and police have independently identified a suspect through other investigative means. Voice line-ups can be constructed (cf. Nolan & Grabe 1996), testing the witness' ability to identify the suspect from a roster of voices. There are contrastive approaches regarding the selection of the foils, specifically whether it should be based on the suspect's voice or the witness' description of it. Proponents of the latter method stipulate that the foils 'meet the speech profile of the suspect […] and be matched for features such as biological and social gender, perceived age, accent and dialect' (Broeders and Van Amelsvoort 2001:241). More recent research advocates the opposing method (cf. Nolan 2003; de Jong-Lendle et al. 2015), that witness statements should not play a role in the construction of the line-up. Indeed, studies have shown that naïve listeners' estimations of age, weight, height, geographical origin, and education are subject to substantial variation in accuracy (cf. Gonzalez 2003, Tomkinson & Watt 2018). What remains to be seen, however, are a) the accuracy of these estimations in relation to each other, because these studies have all been independent, and b) other factors such as sexual orientation or physical size, that may be used by witnesses to describe the voices they heard. Without necessarily supporting either method of foil selection, this study seeks to yield more comprehensive information regarding the ability of lay people to identify social characteristics in voices, an existing knowledge gap that contributes to the controversiality and questioned validity of this forensic application.

More broadly, there are several other factors that affect both approaches, about which we still know relatively little about, and therefore by extension the degree to which they render earwitness evidence error prone and inaccurate (e.g. decay of long-term auditory memory; exposure length and (un)familiarity of speakers etc., cf. Saslove & Yarmey 1980, Stevenage et al. 2011, Öhman et al. 2013). This study also seeks to shed more light on the third issue listed, by investigating any potential correlation between accuracy of dialect identification and extent of dialect awareness.

In this contribution – a proof-of-concept study – we examined how 36 naïve listeners from across German-speaking Switzerland rated a short snippet of speech (one sentence of c. 20 syllables read aloud) from 16 speakers of the SDATS database (Leemann et al. 2020) in terms of age, height, weight, geographical origin, and educational background. The 16 speakers (8F, 8M) were selected so as to be maximally diverse in terms of age, geographical origin, educational background, height, and weight. The 36 listeners (26F, 10M; mean age 31, SD=17) came from the cantons of Aargau, Bern, Luzern,

Graubünden, St. Gallen, Schaffhausen, Solothurn, Schwyz, Wallis, and Zurich. A large proportion of them had a higher education degree (50%), marginally fewer a school-leaving certificate (Matura) (42%), and the remainder a vocational school degree (8%). Each listener was able to play the sentence multiple times via a web-interface (see Fig. 1). Sentences were presented in random order. For age, height, and weight we measured the mean absolute difference (rather than the 'average difference', which would cancel out deviations from true values). For geographical origin, the listeners had to guess the canton of origin. We measured the distance in km between their guess and the true canton of origin, based on the geographic centroids of cantons.

Results revealed the following patterns: for age (Fig. 2), there is an average difference between estimation and true age of 12.8 years. For height and weight (Figs. 3 & 4), the average differences amount to 4.3cm and 7.7kg. For geographical origin (Figs. 5 & 6), the average difference between estimated and true origin is 49.2km. Finally, for educational background (Fig. 7), results showed that for ten out of 16 speakers, education estimations are above or right at chance level (chance level at 25%). For six speakers, guessing the speaker's educational level was below chance.

In terms of a (at this stage impressionistic) ranking of accuracy, then, it appears that estimations of age and geographical origin deserve strong caution, while height and weight guesses are surprisingly more accurate. Note, though, that height and weight guesses are strongly related to the gender of the speaker, which is almost unanimously guessed correctly (and therefore not shown in graphs). The fact that for a majority of speakers, listeners were able to assign educational background above chance is remarkable in the context of Swiss German: typically, it is assumed that variation between speakers is horizontal, i.e. geographically determined – and not vertical (i.e. socially determined – with 'educational background' as a proxy). Perhaps between-speaker variation in reading competence in dialect was a tell-tale sign of the speakers' educational background. In terms of geographical origin, it is conceivable that the inaccuracy of estimations may have to do with the short exposure to the audio samples in the current design. Estimations of geographical origin would likely have been higher, had there been a longer exposure phase.

In future, this study will be expanded to a substantially larger speaker and listener set, using longer exposure times. This will consolidate (or refute) some of the findings of this first pilot study. In terms of ranking of parameters, we will also need to find ways of how the estimations reported can be compared to each other in a more nuanced fashion. This research could form the basis of further study into how auditory memory decay affects the ability to identify region, age, educational background etc., as some parameters may be more resistant than others. In terms of implications for voice line-ups, this study's results indicate that several parameters are less reliable than others when taking witness' testimony into account, namely geographical origin and age.

References

Broeders, A., & van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. Problems of Forensic Sciences, 47, 237-245.

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: a practical guide. In *Proceedings of the International Congress of Phonetic Sciences,* p. 10-14.

Gonzalez, J. (2003). Estimation of speakers' weight and height from speech: A re-analysis of data from multiple studies by Lass and colleagues. Perceptual and motor skills, 96(1), 297-304.

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., Messerli, J. (2020). SDATS Corpus – Swiss German Dialects Across Time and Space. Retrieved from osf.io/s9z4q

Nolan, F. (2003) A recent voice parade. Forensic Linguistics 10(2), 277-291.

Öhman, L., Eriksson, A., & Granhag, P. A. (2013). Angry voices from the past and present: Effects on adults' and children's earwitness memory. Journal of Investigative Psychology and Offender Profiling, 10(1), 57-70.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. Journal of Applied Psychology, 65(1), 111.

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. Applied Cognitive Psychology, 25(1), 112-118.

Tomkinson, J. and Watt, D. (2018). Assessing the abilities of phonetically untrained listeners to determine pitch and speaker accent in unfamiliar voices. Language and Law/Linguagem e Direito, 5(1), 19–37.

Analysing the effect of language on speaker-specific speech rhythm in Cantonese-English bilinguals

Adas Li[1,2], Peter French[1,3], Volker Dellwo[4] and Eleanor Chodroff[1]
[1]University of York, [2]University of Hong Kong, JP French Associates, York,
[4]University of Zürich

Background. Undertaking forensic speaker comparison (FSC) cases in which the questioned sample is in a different language from the known sample has caused concern within the forensic phonetics community (see clause 3.10 of [1]). A major reason for such reservations is the lack of research knowledge concerning which aspects of speech are robust to language switching. To develop an empirically-justified basis for conducting cross-language analysis, it is important to widen the bilingual focus of FSC research. Within a single language, rhythm is known to be a speaker-specific parameter in FSC. However, little is known about its potential as a speaker discriminant for bilinguals in forensic practice. Few studies have examined how rhythm varies between bilingual speakers and between their L1 and L2 languages; and even fewer have examined this between two typologically distinct languages such as Cantonese and English [2][3].

This research extends previous literature on bilingual rhythm variability and L2 rhythm studies to assess the robustness of rhythm characteristics against language-specific effects and to identify speaker-specific parameters within FSC research: 1) Is there significant between-speaker rhythmic variability among Cantonese-English bilinguals?, 2) To what extent do other factors (language and speaker effects) contribute to such variability?, and 3) To what extent can we predict a speaker's rhythm in their L2 English speech based on their L1 rhythmic patterns, and vice versa?

Methods. The speech data used in this study was retrieved from the ALLSSTAR (Archive of L1 & L2 Scripted and Spontaneous Transcripts and Recordings) corpus [4][5]. The speech materials were produced by 14 Cantonese-English bilinguals (8 females and 6 males, mean age = 22, range = 19–27), whose English language proficiency ranged from sufficiently high for US undergraduate studies to native proficiency. All speakers indicated Cantonese as their L1. Speakers produced two 'Hearing in the Noise Test' sentence sets and the 'North Wind and the Sun' passage in both Cantonese and English.

The audio files (44.1kHz sampling rate, 16 bit depth, mono) were segmented by utterance, and then aligned at the phone level with Montreal Forced Aligner [6]. Speech disfluencies were removed and manual correction was carried out to ensure the precision of the alignment. Following [7], the five rhythm measurements (RMs) of rateCV, V%, ΔV(ln), ΔC(ln), and ΔPeak(ln) were taken using the *durationAnalyzer* Praat script [8][9].

Statistical analysis. General observations were made using descriptive analysis and visual plotting in *R*. To investigate the degree to which rhythm varied by language and speaker, a series of linear regression models was performed and compared: a) a model with a single fixed effect of language, b) a Linear Mixed-Effects (LMMs) Model with a fixed effect of language and a random intercept of speaker and, lastly, c) a LMMs Model with a fixed effect of language, a random intercept for each speaker, as well as a random slope of language for each speaker. The three models' goodness of fits were then compared using the Akaike Information Criterion (AIC). To assess the potential utility of speaker-specific rhythm for speaker identification, a closed-set speaker classification task was implemented using a multinomial logistic regression that predicted the speaker's identity from the 5 RMs and their interactions for each utterance. 4 models were trained and tested using mutually-exclusive subsets within and across the 2 languages.

Results. AIC strongly favoured the model with by-speaker intercepts and slopes for language for each of the five RMs. The considerable improvement in model fit indicated sizable speaker-specific influences on overall and language-specific rhythm variability.

The potential of using speaker-specific rhythm for classification was assessed with a multinomial logistic regression. Average accuracy in speaker classification was the highest when the model was trained and tested on the same language: 25% for Cantonese and 19% for English but it decreased in cross-language prediction (in which the model was trained on the rhythm features of one language and tested on those of the other): 12% for the Cantonese-trained-English-tested model and 11% for the English-trained-Cantonese-tested-model. Despite the low accuracy of the overall training models, the accuracy was significantly above chance for all four models, as estimated by bootstrapping with 10,000 samples. This finding reveals some contribution of speaker-specific rhythm for speaker classification both within a language, and critically across the speaker's two languages. Further research is, however, necessary to determine the extent to which classification accuracy is influenced by other intervening variables such as the speakers' level of competence or intelligibility in their L2 English, the type of L1 dialects spoken, as well as the stimuli used in the experiment.

Conclusion. Overall, this paper examines the effect of language and speakers on rhythm variability and finds it of some – but limited – use in speaker comparisons. Consistent with previous findings, we identified significant effects of language and speaker on rhythm variability [2][3], thus leaving reservations about undertaking cross-linguistic analysis within the field of FSC casework relatively unallayed. Nevertheless, a small and significant degree of speaker-specificity is present in the realisation of rhythm. This individuality and cross-language predictability of rhythm could still be useful in FSC cases, at least in combination with more canonical speaker-specific features. These findings highlight the need for all aspects of speech to be researched as potential candidates for inclusion in cross-language comparisons.

References.

[1] International Association for Forensic Phonetics and Acoustics, Code of Practice (2020). http://www.iafpa.net/the-association/code-of-practice/

[2] Dellwo, V., Schmid, S., Leemann, A., Kolly, M.-J., & Müller, M. (2012, July). Speaker identification based on speech rhythm: the case of bilinguals. *Perspectives on Rhythm and Timing (PoRT)*. https://doi.org/10.5167/uzh-111811

[3] White, D., & Mok, P. (2019). L2 Speech Rhythm and Language Experience in New Immigrants. *The 19th International Congress of Phonetic Sciences (ICPhS 2019)*, 1–5.

[4] Ackerman, L., Burchfield, L. A., Hesterberg, L., Bradlow, A. R., Luque, J. S., & Mok, K. (2010). *ALLSSTAR Project Manual*, https://groups.linguistics.northwestern.edu/speech_comm_group/allsstar2/#!/manual

[5] Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://oscaar3.ling.northwestern.edu/ALLSSTARcentral/#!/recordings.

[6] McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner [Computer program]. Version 0.9.0, retrieved 17 January 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[7] Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

[8] Dellwo, V. (2019). *Praat script: Duration Analyzer (version 0.03)*. University of Zurich. https://www.pholab.uzh.ch/static/volker/software/plugin_duratio%0AnAnalyzer.zip

[9] Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.22, retrieved 24 September 2020 from http://www.praat.org/.

Seeing the trees in the forest: Diagnosing individual performance in likelihood ratio based forensic voice comparison

Justin J. H. Lo (University of York, UK)
jl2355@york.ac.uk

Introduction

In numerical likelihood ratio (LR) based forensic voice comparison (FVC), system testing is predominantly evaluated on the global level. System validity is assessed by means of a single metric score representing the proportion or size of errors, the most commonly used of which are the equal error rate (EER) and log likelihood ratio cost function (C_{llr}). Other graphical means of illustration, such as ROC curves, DET curves and Tippett plots, are also regularly used to provide more information about the overall performance of the system.

As much as such measures and graphs can provide an indication of global system performance, their diagnostic value can be limited. Within any given system, variation of performance between individual speakers may arise due to physiological, behavioural and technical reasons. Analysis of performance on the level of individual speakers thus offers the potential to gain insights into the nature of errors and, as such, for work towards improving system design in a targeted manner. In addition, an examination of individual performance in relation to the acoustic-phonetic data used to generate output LRs can further our understanding of the relationship between input and output in LR based FVC.

As analysis of individual performance remains rarely performed in the context of FVC, the present study seeks to demonstrate its utility in LR based FVC and further explores the connection between individual performance as derived from LRs and the underlying speech data. As a case study, this study makes use of long-term formant distributions (LTFDs), a set of features argued to be able to capture both anatomical variation of the vocal tract and idiosyncratic articulatory habits of speakers.

Methods

The present study is part of an ongoing project that draws its data from the Voice ID Database (RCMP 2010–2016). High-quality microphone recordings from 60 adult male bilingual speakers of Canadian English and French were analysed, but the current exploratory analysis is limited to the English materials, consisting of phonetically balanced read sentences and passages. All recordings were automatically segmented using the Montreal Forced Aligner (McAuliffe et al. 2017), followed by manual checks and corrections where necessary. Formant estimates for the first four formants were automatically extracted in Praat at 10ms intervals from all vowels and glides, with the formant tracker set to search for 6 formants up to a maximum formant frequency of 5500 Hz in 25ms frames.

To evaluate the performance of LTFDs as speaker discriminants, speakers were randomly partitioned into three equally sized of test, training and reference speakers. LTFDs from all four formants combined were modelled and compared using GMM–UBMs (Reynolds et al. 2000) to generate scores, which were then calibrated by logistic regression to obtain $\log_{10}$LRs (LLRs). LTFDs from each formant were likewise tested separately to facilitate one-to-one comparison with the acoustic data. C_{llr} and EER were calculated to assess the overall performance of each system. The sampling and testing procedure was replicated 100 times, in order to minimise the effects of random speaker sampling on the comparison of LLRs and metrics of validity.

For analysis on the individual level, this study makes use of the notion of the "biometric menagerie" (Doddington et al. 1998), where speakers are classified into user

groups, or animals, based on their performance. Zooplots were constructed by plotting a speaker's average performance in different-speaker comparisons against their performance in same-speaker comparisons. Average performance of any speaker in this study is defined as the arithmetic mean of LLRs from all same- or different-speaker comparisons (SS-LLR; DS-LLR) involving that speaker. Following the definitions in Dunstone and Yager (2009), speakers whose performance ranked within the top or bottom quartile of all speakers, in both same-speaker and different-speaker comparisons, were identified and classified as doves, worms, phantoms and chameleons, and their LTFD data were further analysed with respect to other speakers in the group.

Results and discussion

In terms of global system performance, all systems using LTFDs from single formants reported similar mean C_{llr} (0.62-0.69) and EER (19-22%), suggesting that overall each LTFD performed at a similar level. The system combining LTF1-4 performed considerably better, as evidenced by the lower mean C_{llr} (0.31) and EER (7%).

Zooplots show that, for each formant, all speakers reported a negative mean DS-LLR, indicating that, on average, they were capable of being distinguished from other speakers. While the majority of speakers had a positive mean SS-LLR, some speakers reported negative mean SS-LLRs, indicating same-speaker comparisons as a source of errors.

Zoo analysis further shows that, among the set of 60 speakers, different subsets of speakers were identified as doves and worms for each individual formant, thus providing corroborating evidence on the individual level that each LTFD captures some complementary speaker-specific information. Comparison with the underlying acoustic data shows clear separation between doves and worms, especially in LTF3 and LTF4. The distributions of doves typically exhibited peaks at more extreme frequencies, while those of worms had similar shapes to the overall distribution of the group.

The findings above demonstrate the diagnostic value that individual-level analysis can add to system evaluation in LR based FVC by providing a more fine-grained picture of performance. Further analysis of the acoustic-phonetic data illustrates how individual distributions of acoustic-phonetic data can be reflected in exceptional speaker-discriminatory performance. This study thus supports an approach where LR based FVC is concerned with not only global measures of validity and reliability, but also the performance of individual speakers and the factors behind their performance.

References

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 Speaker Recognition Evaluation.

Dunstone, T. & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. New York: Springer.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). *Montreal Forced Aligner* [computer programme]. Version 1.0.0, retrieved 30 November 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12(2), 143–173.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 19–41.

Royal Canadian Mounted Police [RCMP]. (2010–2016). Voice ID Database [unpublished audio corpus]. Collected at University of Ottawa.

I silenzi dei matti. Gli spazi 'vuoti' del parlato nell'archivio sonoro di Anna Maria Bruzzone

Rosalba Nodari, Silvia Calamai - Università di Siena

Introduzione

Un filone di studi, nato principalmente in seno alla linguistica applicata e all'analisi del discorso, rivolge la sua attenzione alla verbalizzazione di quella che può essere considerata l'esperienza traumatica. Secondo i paradigmi teorici delle filosofie *embodied*, l'esperienza del trauma lascia infatti tracce visibili che emergono nel momento in cui chi è stato vittima di esperienza traumatica è chiamato a narrare e, quindi, a compiere una costruzione discorsiva del trauma come un oggetto di conoscenza (Ogden & Minton, 2000; Busch, 2017; Busch & McNamara, 2020). Lavori condotti sulle narrazioni di esperienze traumatiche hanno quindi cercato di cogliere gli elementi linguistici ricorrenti, le strategie retoriche e conversazionali, gli stili narrativi, le prese di turno, e hanno fatto emergere come i parlanti reagiscano all'esperienza in maniera estremamente peculiare (per una retrospettiva di studi, cf. Busch & McNamara, 2020 e i lavori contenuti nella *special issue*). Purtuttavia, fino ad ora gli studiosi hanno principalmente rivolto il proprio sguardo sugli spazi 'pieni' del discorso, tralasciando invece gli spazi 'vuoti', le lacune e le assenze. Eppure, spesso non attraverso le parole, bensì attraverso i silenzi è possibile esprimere la sostanza del trauma (Harjula, 2002): l'assenza della verbalizzazione è spesso la conseguenza di una negazione del diritto di parola e dell'essere riconosciuti come soggetto del discorso; proprio in questi silenzi si può manifestare la rottura del sé (Ferenczi, 1916; Ritter, 2014; Busch, 2017). L'idea del silenzio come metafora per la comunicazione, con il silenzio elemento fondamentale – al pari del parlato – per l'interpretazione di una frase, permette di superare l'equazione silenzio = assenza di suono, e di individuare in esso una possibile strategia per la manifestazione di fenomeni plurimi – linguistici, pragmatici, socioculturali, metacomunicativi (Jaworsky, 1977; Tannen & Saville-Troike, 1985).

In parallelo agli studi sul cosiddetto *trauma speech,* ricerche condotte in ambito clinico hanno preso in considerazione elementi linguistici, che possono essere visti come un buon indice per classificare alcune patologie psichiatriche come la schizofrenia o la depressione. Ricerche in questo ambito si sono concentrate non solo su aspetti discorsivi e lessicali, ma anche su caratteristiche acustiche, quali il range dinamico della frequenza fondamentale, la velocità d'eloquio, le caratteristiche prosodico-intonative, e seppur in maniera minore, la quantità di pause silenti e di elementi non verbali (Tolkmitt et al., 1982; Low et al., 2010; Cummins et al., 2011; Xu et al., 2018, 2019; Parola et al., 2020). Stando alla conoscenza di chi scrive, solo Salah et al. (2019) hanno tentato un primo incontro tra le osservazioni di natura acustica relative al parlato patologico e al parlato emotivo collegato all'esperienza di tipo traumatico, con un particolare focus sui silenzi e sui momenti di respiro. In particolare, Salah et al. (2019) offrono una prima analisi del parlato di 10 parlanti sopravvissuti a diversi eventi traumatici di massa, confrontando brani di parlato emotivamente più neutri con brani con narrazioni di eventi traumatici. L'analisi mette in evidenza che, pur con le dovute differenze culturali, emergono delle costanti relativamente alle fasi di profondo respiro e ai silenzi.

Sulla base di Salah et al. (2019), il seguente lavoro vuole offrire una prima analisi dei momenti di silenzio e di respiro profondo di un particolare gruppo di parlanti, ossia i degenti dell'ospedale psichiatrico di Arezzo, oggi Campus del Pionta, intervistati dalla storica Anna Maria Bruzzone.

Il corpus

Nel 2016 è stato ritrovato, a Torino, l'archivio sonoro alla base del volume *Ci chiamavano matti. Voci da un ospedale psichiatrico* (Einaudi, Torino 1979) di Anna Maria Bruzzone: 24 audiocassette con 16 voci maschili e 18 voci femminili, con associate le trascrizioni originali in differenti versioni. L'archivio sonoro è stato digitalizzato grazie al sostegno della Soprintendenza Archivistica e Bibliografica della Toscana ed è al momento in corso di catalogazione e di analisi (Calamai e Biliotti, 2017). Solo grazie al ritrovamento delle registrazioni è stato possibile portare a termine studi che prendessero in considerazione la 'viva voce dei matti', svincolandosi dai limiti imposti dalla trascrizione normalizzante compiuta dalla stessa Bruzzone e necessaria per la pubblicazione del volume. L'archivio sonoro permette inoltre l'analisi di quello che per sua stessa natura non può essere reso nello scritto: i respiri profondi, le pause e i silenzi degli intervistati. L'analisi dei silenzi di questo tipo di materiale risulta inoltre fondamentale per la stessa natura epistemologica alla base della ricerca di Bruzzone. Intenzione della storica era quella di dare voce a chi, fino a poco fa, era stato costretto a tacere (Vangelisti et al., 2019). I parlanti di Bruzzone recuperano quindi la parola in un momento peculiare in cui, grazie all'esperienza basagliana di apertura del manicomio, diventano protagonisti grazie anche ai numerosi momenti di confronto durante le assemblee generali e dei reparti.

La metodologia

Il lavoro vuole offrire un modello che permetta di analizzare nel dettaglio il rapporto tra spazi 'pieni' e 'vuoti' del parlato, intendendo con questi ultimi le pause, i silenzi, i respiri. Il nostro metodo di analisi verrà poi applicato a una selezione dei 34 parlanti intervistati da Bruzzone nel 1977. A questi aggiungeremo uno studio di caso dedicato alle interviste condotte con R., ex degente intervistato da Bruzzone, rintracciato nel 2017 e nuovamente intervistato da Caterina Pesce, per la sua tesi dottorale (Vangelisti et al., 2019).

Le interviste sono state trascritte dalle AA. Le trascrizioni così generate sono state poi rese come textgrid grazie alla procedura di forced alignment offerta da WebMAUS (Kisler et al., 2017) e importate in PRAAT 6.0.36 (Boersma & Weenink, 2015).

All'analisi strumentale tramite Praat viene affiancata un'analisi compiuta con il programma Voyant (voyant.org), che permette di rendere conto della densità lessicale delle interviste. Il modello mira a offrire una prima classificazione globale del parlante prendendo in analisi l'intervista nella sua interezza. I parametri tenuti in considerazione sono: i) un indice medio di 'ricchezza lessicale' del parlante, calcolato come il rapporto tra numero di parole/durata dell'intervista ii) un indice relativo alla velocità d'eloquio globale del parlante, calcolata sia come articulation rate all'interno di una sequenza di parlato connesso (sillabe/sec), sia come *speaking* rate, includendo cioè anche le pause; iii) una durata cumulativa delle pause del parlante, calcolata come il rapporto tra durata dei silenzi all'interno di un turno conversazionale / durata dell'intervista. In aggiunta alla classificazione globale, si vuole osservare come variano i parametri sopra elencati in determinate stringhe di parlato in cui si revoca l'esperienza traumatica. Per effettuare un raffronto tra momenti in cui il *topic* del discorso era chiaramente legato all'esperienza traumatica della degenza in manicomio e altri momenti emotivamente più neutri si è condotta un'analisi lessicale sempre attraverso Voyant.

Per l'individuazione degli spazi vuoti, terremo in considerazione solo le pause, definite da Sacks, Schegloff & Jeffeson (1974) come i momenti di silenzio che compaiono all'interno di un turno conversazionale prodotto da uno stesso parlante senza interruzioni (cf. Ten Bosch, Oostdijk & Boves 2005; Heldner & Edlund 2010). Attraverso questa definizione potremo così distinguere tra pause (*turn internal*) e *gaps* (silenzi tra parlanti). Per evitare di includere nell'analisi discontinuità dovute a fenomeni articolatori, come ad es. la fase di chiusura delle occlusive (Goldman-Eisler 1968: 12) considereremo come pause utili all'analisi solo quelle di durata superiore a 100 ms. La decisione di questo valore soglia è stata fatta in accordo con la bibliografia fonetica consultata (Campione & Veronis 2002; Clopper & Smiljanic 2015; Heldner & Edlund 2010; Kendall 2009, 2013; Ten Bosch et al. 2005).

I dati presentati al Convegno (per ora relativi solo alle pause in ms e ancora in corso di lavorazione), vogliono anche dimostrare come l'analisi strumentale di tipo acustico possa essere di supporto non solo per analisi di tipo medico cliniche, ma anche per quegli studi che vedono al centro la costruzione narrativa del sé, che di solito vedono l'applicazione di criteri di trascrizione ascrivibili all'analisi conversazionale. Una analisi di tipo acustico può al contrario rappresentare un modo per rendere visibile e quantificabile ciò che vive dentro e attraverso la voce dei parlanti.

Bibliografia

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer* [computer program](2011). Version, 5(3), 74.

Bruzzone, A. M. (1979). *Ci chiamavano matti. Voci da un ospedale psichiatrico*. Torino: Einaudi.

Busch, B. (2017). Expanding the notion of the linguistic repertoire: On the concept of Spracherleben – the lived experience of language. *Applied Linguistics,* 38: 340–58.

Busch, B., & Mcnamara, T. (2020). Language and Trauma: An Introduction. *Applied Linguistics,* 41(3): 323-333.

Calamai, S. & Biliotti, F. (2017). Le voci dei matti. Il ritrovamento dell'archivio sonoro di Anna Maria Bruzzone. In Baioni, M., Setaro, M. (Eds.), *Asili della follia. Storie e pratiche di liberazione nei manicomi toscani*. Pisa: Pacini editore, pp. 22-34.

Campione, E. & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 Conference* (pp. 199-202). Aix en Provence: Laboratoire Parole et. Langage

Clopper, C. G., & Smiljanic, R. (2015). Regional variation in temporal organization in American English. *Journal of Phonetics*, 49: 1-15.

Cummins, N., J. Epps, M. Breakspear and R. Goecke. (2011). An investigation of depressed speech detection: Features and normalization. *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Ferenczi, S. (1916). Silence is golden. In *Further contributions to the theory and technique of psychoanalysis*. London: Karnac, 1994, pp. 250–251.

Harjula, E. (2002) Trauma Lives in Speech. *International Forum of Psychoanalysis*, 11(3): 198-201.

Heldner, M. & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics,* 38.4: 555-568.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* New York: Academic Press.

Jaworski, A. (Ed.) (1997). *Silence: Interdisciplinary Perspectives*. Berlin and New York: Mouton de Gruyter.

Kendall, T. (2009). *Speech Rate, Pause, and Linguistic Variation: An Examination Through the Sociolinguistic Archive and Analysis Project*. PhD Thesis: Duke University.

Kendall, T. (2013). *Speech rate, pause and sociolinguistic variation: studies in corpus sociophonetics*. Berlin: Springer.

Kisler, T., Reichel, U. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language* 45(2017): 326-347.

Low, L.-S. A., N. C. Maddage, M. Lech, L. Sheeber & N. Allen. (2010). Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. *Acoustics Speech and Signal Processing* (ICASSP), 2010 IEEE International Conference, pp. 5154–5157, IEEE, 2010.

Ogden, P. & Minton, K. (2000). Sensorimotor psychotherapy: One method for processing traumatic memory. *Traumatology,* 6(3): 149.

Parola, A., Simonsen, A., Bliksted, V. & Fusaroli, R. (2020). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizophrenia Research,* 216: 24-40,

Ritter, M. (2014). Silence as the Voice of Trauma. *The American Journal of Psychoanalysis*, 74(2): 176-194.

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4): 696-735.

Salah, Aa A., Ocak, M., Kaya, H. Kavcar, E. (2019). Hidden in a Breath: Tracing the breathing patterns of survivors of traumatic events. Abstract of paper 0982 presented at the *Digital Humanities Conference 2019* (DH2019), Utrecht, the Netherlands 9-12 July, 2019.

Tannen, D. & Saville-Troike, M. (Eds.) (1985). *Perspectives on Silence*. Norwood, NJ: Ablex.

Ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2): 80-86.

Tolkmitt, F., Helfrich, H., Standke, R., & Scherer, K. R. (1982). Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of communication disorders*: 15(3), 209-222.

Vangelisti, P., Pesce, C., Setaro, M., Bianchini, G., Gigli, L., & Calamai, S. (2019). Ritrovare Voci: il lavoro intorno all'archivio di Anna Maria Bruzzone. In Piccardi, D., Ardolino, F. & Calamai, S. (Eds.) *Gli archivi sonori al crocevia tra scienze fonetiche, informatica umanistica e patrimonio digitale - Audio archives at the crossroads of speech sciences, digital humanities and digital heritage.* Milano: Officinaventuno, pp. 155-168.

Xu, S., Yang, Z., Chakraborty, D., Tahir, Y., Maszczyk, T., Chua, V. Y. H., ... & Keong, J. L. C. (2018). Automatic verbal analysis of interviews with schizophrenic patients. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)* (pp. 1-5). IEEE.

Xu, S., Yang, Z., Chakraborty, D., Chua, Y. H. V., Dauwels, J., Thalmann, D., ... & Keong, J. L. C. (2019,). Automated Verbal and Non-verbal Speech Analysis of Interviews of Individuals with Schizophrenia and Depression. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 225-228). IEEE.

Discriminating speakers using perceptual clustering interface

Benjamin O'Brien[1], Alain Ghio[1], Corinne Fredouille[2], Jean-François Bonastre[2], Christine Meunier[1]

1 Aix-Marseille Univ., CNRS, LPL, UMR 7309, Aix-en-Provence, FR

2 Laboratoire d'Informatique d'Avignon, Avignon Université, Avignon, FR

INTRODUCTION The challenges facing listeners tasked to identify speakers are well documented.[1 2 3] In addition to providing listeners with high-quality speech recordings that accurately represent the speakers, the method of presentation itself is equally important.[4 5] Numerous perception studies have employed a binary approach, where participants are asked to judge whether two speech recordings are similar or different, as a way of examining the effects of such things as noise,[6] language familiarity,[7 8] and stimuli selection methods.[9] Oftentimes this requires numerous tests, which can be time-consuming for participants and experimenters. Moreover, there persists concern for memory bias, as a "fresh" voice is not equivalent to a voice that was presented in a previous binary test.

As an alternative, we proposed the development of a perceptual *clustering* method, which is often employed in the domain of machine learning.[10 11] We theorized that this approach would allow users to better personalize their engagements with speech materials and organize their proximities in relation to their perceived likeness. In addition, it was more economical in terms of the number of trials required to assess a listener's ability to identify speakers.

In order to study the speaker discrimination performance of participants using a perceptual clustering interface, it was important to organize and select stimuli based on how listeners perceive them as similar or different. Studies suggest listeners rely on a common set of acoustic features to identify speakers.[12 13] It is common in the development of automatic voice recognition and speaker identification system to extract MFCCs from speech recordings to train models. A popular trend in the field involves the transformation of these features into i-vectors, which have been shown to be quite accurate in identifying speakers.[14 15] Recent work has shown that Cosine Distance Scoring (CDS) with Within-Class covariance normalization (WCCM) is similarly effective while reducing the complexity of the task.[16] Our second objective was to examine the relationships between participant performance and the CDS generated from the speaker i-vectors.

METHODS Speech recordings were selected from the PTSVox database,[17] which included 24 francophone speakers (12 female, 12 male) who recited three French-texts into a Zoom H4N stereo microphone (sampling rate: 44.1 kHz; bit depth: 16-bit) over the course of two recording sessions (mean 118.96 ± 17.54 s). SPro[18] was used to extract 19 MFCCs, deltas, and delta-deltas from each recording. ALIZE[19] was used to compress these features into i-vectors and then calculate CDS between each one, whereupon the WCCM was computed over the entire set. Two groups of five speakers were selected: the *Alpha* group was composed of speakers with the greatest distance between them and the *Betha* group was composed of speakers with the smallest distance between them. For each speaker, twelve utterances were selected (120 recordings; mean 1.47 ± 0.51 s). Groups were divided into six sessions, such that each session was balanced and composed of four different (non-repeating) chunks per speaker.

Twenty-four people (14 female; 24.2 ± 6.7 years), who self-reported good hearing, participated in our study. Their task was to group 20 speech recordings into five cluster groups, where each cluster represented a unique speaker. To do this, they used the TCL-LABX interface,[20] which allowed them to move recordings in a 2-D space and assign them to different clusters. They completed six sessions.

The Mathews Correlation Coefficient (MCC) was selected to determine how accurate the participants were at discriminating speakers (1), where *TP, TN, FP, FN* represent the selections that were "true positive," "true negative," "false positive," and "false negative," respectively. The mode speaker in each cluster was used to calculate the MCC. MCC means and standard deviations were calculated for each speaker.

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))}} \quad (1)$$

RESULTS To examine participant performance discriminating speakers, two-level nested ANOVA procedures were applied to MCC mean and standard deviation for groups with different speakers. We found a main effect on groups for MCC mean $F_{1,240} = 32.92$, $p < 0.001$, $\eta_p^2 = 0.12$, and no significant differences between speakers within each group, $p > 0.05$. Post-hoc tests revealed Alpha had a higher MCC mean (0.94 ± 0.20) when compared to Betha (0.8 ± 0.02), $p < 0.001$. Similarly we found a main effect on MCC standard deviation $F_{1,240} = 26.04$, $p < 0.001$, $\eta_p^2 = 0.1$, but again no significant differences between speakers

within each group, $p > 0.05$. Post-hoc tests revealed Alpha had a lower MCC standard deviation (0.08 ± 0.02) when compared to Betha (0.2 ± 0.02), $p < 0.001$ (Fig. 1a).

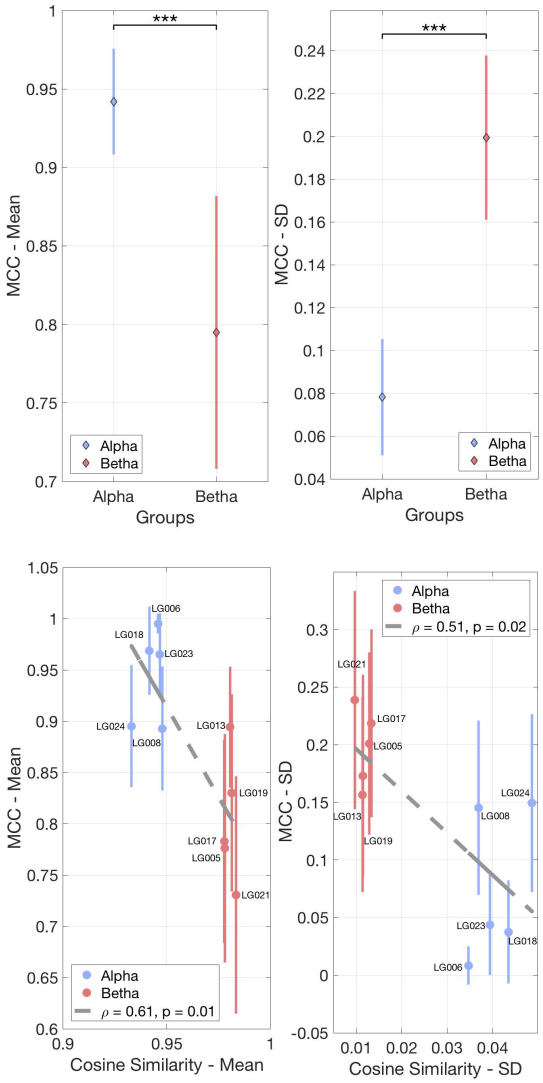


Figure 1a: Mean (Left) and standard deviation (Right) of participant MCC per group. Diamonds and vertical lines represent the means and standard errors, respectively. {***} signifies $p < 0.001$. **Figure 1b**: Pearson's coefficient was used to examine the relationship between CDS and MCC metrics: mean (Left) $\rho = 0.61$, $p = 0.01$, and standard deviation (Right) $\rho = 0.51$, $p = 0.02$. Circles and vertical lines represent the means and standard errors, respectively. The text indicates the speaker id.

We then examined whether our method of selecting and grouping speakers played a role in participant performance. For each speaker we calculated the CDS mean and standard deviation between it and the other group speakers and then calculated the Pearson's correlation coefficient to examine the relationships between the two metrics. The speaker CDS mean difference estimated the MCC mean at $\rho = 0.61$, $p = 0.01$, whereas the speaker CDS standard deviation estimated the MCC standard deviation at $\rho = 0.51$, $p = 0.02$ (Fig. 1-b).

DISCUSSION This study demonstrated that users were able to use a clustering interface to make discriminations based on their perceived differences between speech recordings. Participants performed at a relatively high level, as indicated by the mean and standard MCC values, which suggests they found the interface easy to navigate and efficient to use. In addition, the significant differences between groups also underscore the importance of developing methods for selecting and grouping speakers. We observed that as the CDS mean increased, participants were less accurate discriminating speakers, and, conversely, as the CDS standard deviation decreased, participants showed greater variability. These findings have led us to develop a new study to compare the effects of presentation on users performing speaker discrimination tasks with similar speaker stimuli.

References

[1] Cambier-Langeveld, T., Rossum, M., Vermeulen, J. (2014). Whose voice is that? Challenges in forensic phonetics.

[2] Mattys, S. Davis, M., Bradlow, A., Scott, S. (2012). Speech recognition in adverse conditions: A review. Language and Cognitive Processes - LANG COGNITIVE PROCESS. 27. 953-978.

[3] Nolan, F. (2001). "Speaker identification evidence: its forms, limitations, and roles," Law and Language: Prospect and Retrospect.

[4] Boë, L., Bonastre, J-F. (2012). L'identification du locuteur: 20 ans de témoignage dans les cours de Justice. Le cas du LIPSADON « laboratoire indépendant de police scientifique », JEP, Grenoble: 417-424.

[5] Hollien, H., Bahr, R., Künzel, H., Hollien, P. (2013). "Criteria for earwitness lineups," Int. Jnl of Speech Language and the Law 2, 143–153.

[6] Smith, H., Baguley, T., Robson, J., Dunn, A., Stacey, P. (2018). Forensic voice discrimination: The effect of speech type and background noise on performance. Applied Cognitive Psychology.

[7] Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language familiarity effect for speaker discrimination without comprehension. Proc National Academy of Sciences, 111(38)

[8] Levi, S. V., & Schwartz, R. G. (2013). The development of language specific and language-independent talker processing. Journal of Speech, Language, and Hearing Research, 56(3), 913–920.

[9] Mühl, C., Sheil, O., Jarutytė, L. et al. (2018) The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. Behav Res 50, 2184–2192.

[10] Kinnunen, T. and Kilpeläinen, T. (2000). Comparison of clustering algorithms in speaker identification. Proc. IASTED Int. Conf. Signal Processing and Communications (SPC). 222-227.

[11] Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T. (2016). Speaker identification and clustering using convolutional neural networks. 1-6.

[12] LaRiviere, C. (1971). "Some acoustic and perceptual correlates of speaker identification," Proc 7th Int. Congress Phonetic Sciences: 558-564.

[13] Roebuck, R., and Wilding, J. (1993). "Effects of vowel variety and sample length on identification of a speaker in a line-up," *Applied Cognitive Psychology* 7: 475–481.

[14] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. (2011) Front-End Factor Analysis for Speaker Verification, in IEEE Transactions on Audio, Speech, and Language Processing, 19(4): 788–798.

[15] Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M. (2011). i-vector Based Speaker Recognition on Short Utterances. Proc International Speech Communication Association, INTERSPEECH.

[16] Fredouille, C., Charlet, D. (2014) Analysis of I-Vector framework for Speaker Identification in TV-shows. Interspeech, Singapour, Singapore.

[17] Chanclu, A., Georgeton, L., Fredouille, C., Bonastre, J-F. (2020) PTSVOX: une base de données pour la comparaison de voix dans le cadre judiciaire. 6e conférence conjointe Journées d'Études sur la Parole, Nancy, FR. pp.73-81

[18] Speech Signal Processing (SPro) Toolkit. https://www.irisa.fr/metiss/guig/spro

[19] Larcher, A., Bonastre, J-F., Fauve, B, et al. (2013) "ALIZE 3.0 - Open source toolkit for state-of-the-art speaker recognition," Interspeech, Lyon.

[20] Gaillard, P. (2009). Laissez-nous trier ! TCL-LabX et les tâches de catégorisation libre de sons.

Clustering of unknown voices

Hanna Ruch, Andrea Fröhlich & Martin Lory, Zurich Forensic Science Institute

Introduction

A common request in a forensic phonetician's routine may be the following: How many different speakers do the recordings contain? This question may arise, for instance, in cases where the police wiretapped telephone lines to record phone frauds such as the so-called grandparent scam.

The question above can be addressed using the auditory-phonetic approach, by which the different voices on the recordings are analysed auditorily based on a protocol. Following this approach the voices are analysed and described for voice quality, segmental and supra-segmental features, para-linguistic characteristics, pausing behaviour, and syntax, among others. Features are normally compared against a standard variety, a spoken norm, or a neutral voice, and are made note of on a protocol. Voice descriptions are then compared against each other, and voices which share the same or a major number of relevant features are grouped together. A further possibility to cluster voices is using the phonetic-acoustic approach, which was not conducted for the present dataset.

This procedure becomes very time-consuming and expensive in cases with a large number of recordings. In these cases, approaches based on automatic speaker comparison and (statistical) cluster analysis can be an appropriate alternative.

In this paper we present such an approach based on the automatic speaker comparison system VOCALISE (Kelly et al 2019) and a number of different cluster analysis methods. The modelling technique of the newest version of VOCALISE, called xVocalise, is based on deep neural networks, an advanced machine learning technique (for details see Kelly et al 2019). The procedure described in this paper was applied to a real case in which seven voices on five different recordings had to be grouped according to their similarity.

Method

In a first step an auditory-phonetic analysis and auditory clustering were carried out for the seven voice samples. The files were then pre-processed in such a way that each interval contained only one (hypothetically) different voice. Clipping and non-human noise such as ring tones were removed.

The pre-processed (cleaned) recordings were then analysed automatically using VOCALISE (version 2019A-XVector) based on x-vectors, which is the current state-of-the-art approach in automatic speaker comparison. The comparison between all possible recording pairs was conducted using Linear Discriminant Analysis (LDA) and cosine distance (for details see Kelly et al. 2019) and resulted in a cosine similarity matrix.

Statistical analysis was conducted using R (R Development Core Team 2020). The following clustering methods were applied to the similarity matrix:

- Clustered heatmap
- Multidimensional scaling (MDS) and k-means clustering

Additionally, the x-vectors (a high-dimensional representation of each recording, i.e. each voice) were used as input to the following clustering methods:

- Principal Component Analysis (PCA)
- t-distributed stochastic neighbour embedding (t-SNE)

Results

The statistical methods generally confirmed the results of the auditory analysis (three different speakers). As an example, Figure 1 shows the results of the clustered heatmap. The three blueish clusters indicate that there are three groups of similar voices. According to the visualisation, Speaker3 is more similar to Speaker2 than to Speaker1. This result reflects the auditory analysis, where Speaker2 and Speaker3 were also perceived to be more similar.

The results of the different clustering methods also showed that variation within one hypothetical speaker can be considerable. A closer look at the data suggested that not only recording quality but also the communicative situation and dialogue partner affects the result of the automatic system and, ultimately, the goodness of fit of the clustering.

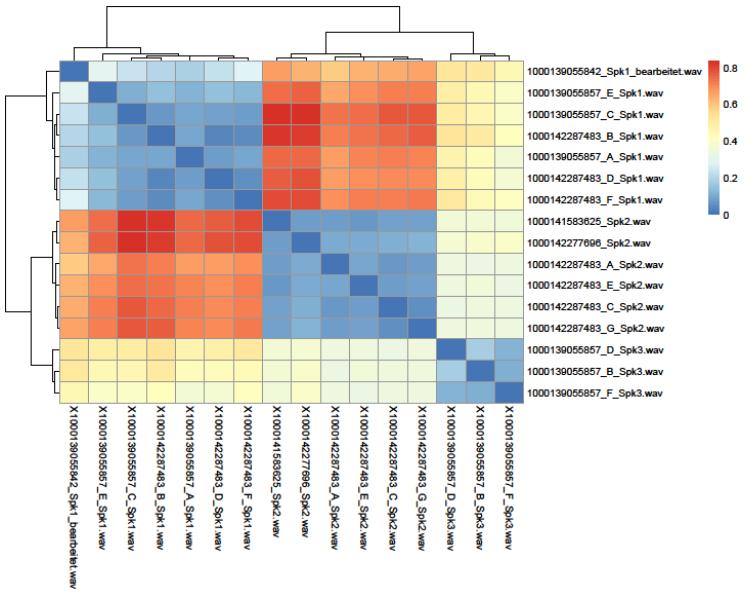


Figure 1: Clustered heatmap showing the cosine distance between all recording pairs. Blueish colours indicate similarity, reddish colours indicate dissimilarity between pairs of voices. The labels Spk1, Spk2, Spk3 stand for the result of the auditory clustering; A-F stand for different intervals (voice samples) within a recording in which two questioned speakers alternated.

Discussion and future directions

For the present dataset and the concrete application in casework the clustered heatmap appeared to be the most appropriate method because it is less abstract and can be explained in a more straight-forward way than the other techniques.

Seven different voices within a dataset is still a rather manageable number, which in could easily be analysed using the auditory approach. Given that the speakers in the dataset are unknown, the results of the two approaches - auditory and automatic/statistical approach - a method validation was not possible. In an on-going project we are currently improving, testing, and validating the methods on a larger, non-forensic dataset with telephone recordings of known speakers. Results are expected to be ready for the AISV conference.

References

Kelly, Finnian, Oscar Forth, Samuel Kent, Linda Gerlach, and Anil Alexander (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors, Audio Engineering Society (AES) Forensics Conference 2019, Porto, Portugal.

R Development Core Team (2020). R: A language and environment for statistical computing. URL: http//www.r-project.org

Prosodic marking of information status in L1 Italian and L2 German

Simona Sbranna, Caterina Ventura, Aviad Albert, Martine Grice
Universität zu Köln

Keywords*: L2 prosody, L2 intonation, periogram, Givenness.*

Introduction. Previous studies on prosodic marking of information status within noun phrases have found that unlike West-Germanic languages, in which post-focal given information is deaccented, in Italian the second word of a noun phrase is always accented independently of information status (Swerts, et al., 2002). Italian learners transfer this feature to their L2 productions (Avesani, et al., 2015). However, the categorical description offered by these studies does not provide information about modulation of continuous prosodic parameters. To address this gap, we test a new method to investigate how far prosody is modulated to mark information status in both native Italian (L1) and Italian learners of German (L2) at beginner, intermediate and advanced proficiency levels (Council of Europe, 2001).

Method. We designed two comparable versions of a semi-spontaneous board game eliciting three information structures – new-new (NN), given-new (GN), new-given (NG) – within noun phrases (NPs) with a disyllabic paroxytone noun and adjective (eg.: *mela nera*; *grüne Welle*). The NPs always occurred in the same syntactic and pragmatic contexts and had the same degree of contextual expectedness, as they were visually represented in the game instructions. We use periograms to display F0 trajectories as modulated by periodic energy (Albert, *et al.*, 2018, Albert, *et al.*, 2020) and extract measures relative to periodic energy, i.e. synchrony and scaling (Cangemi, *et al.*, 2019). Synchrony is the distance between the *Centre of Gravity* (CoG) of F0 (Barnes *et al*., 2012) and the *Centre of Mass* (CoM) of periodic energy within a syllable. Its negative and positive values respectively indicate falling or rising F0 *within* that syllable. *Scaling* is obtained by comparing F0 values at the CoM between two consecutive syllables. Its negative and positive values respectively describe falling or rising F0 movement *across* those syllables.

Results. We report on preliminary results for 15 speakers, five for each level of proficiency. Figure 1 provides examples of periograms and periodic energy curves for GN and NG noun phrases by the same advanced speaker in L1 and L2. In both languages the timing of the falling F0 is distinct across the two conditions. For **L1** in **GN** condition, positive synchrony values on the first syllable indicate a rising F0 on this syllable and positive scaling values on the second syllable indicates a rise across the first and second syllables. By contrast, in **NG** condition the negative synchrony value on the first syllable indicates a predominantly falling F0 and the negative scaling value on the second syllable indicates a fall across the first and second syllables. For **L2**, the **GN** condition has positive values for synchrony on the first syllable, indicating rising F0, but zero scaling value for the second syllable, indicating a stable F0 across the two syllables. In **NG**, the values for both synchrony and scaling are similar to L1, in that they are negative. In all cases, the third syllable has the lowest F0 with the most negative values for both synchrony (F0 is mostly falling within this syllable) and scaling (F0 is mostly falling from the second syllable to the third). The values averaged across all speakers for synchrony and scaling [Fig.2] confirm that these observations are representative. The main difference between the two patterns is achieved by modulation of F0 on the first word in the NP rather than on the second, where it might be expected, as it is there that deaccentuation occurs in German. Although **L2** patterns are similar to those in **L1**, the wide distribution of synchrony values on the first syllable indicates less control over the early modulation. Moreover, the fact that synchrony and scaling values in the second and third syllables are closer to zero in L2 German compared with L1 Italian shows a weaker differentiation in L2 and perhaps also diminished expressiveness. Differences between L1 and L2 are also confirmed by a preliminary analysis of the variance in the distribution of values. Results of Levene's test of homogeneity of variance (Levene, 1960) registered for synchrony a higher variance in L2 than in L1 ($F(1) = 25.51$, $p < .0001$), indicating less control over the F0 modulation in L2, whereas for scaling a higher variance in L1 than in L2 ($F(1) = 13.8$, $p < .001$), indicating an overall reduced F0 range in L2. In L2 we found an effect of proficiency on synchrony value distribution, with advanced learners showing less variance than intermediate ($F(1) = 7.42$, $p = .01$) and beginners ($F(1) = 8.22$, $p < .01$), indicating more coherence in modulating F0 within syllables.

Conclusion. Analysing continuous prosodic parameters (F0 and periodic energy and related measures) is especially beneficial when analysing dynamic and complex systems like interlanguages, which are often unsuitable for a purely categorical analysis. This method allowed us to track the subtle modulations of prosodic cues used to mark different information structures in both L1 and L2. The analysis of the whole corpus, as well as a comparison with L1 German data, which is currently being recorded, will shed more light on the acquisition of strategies for prosodic marking of information status.

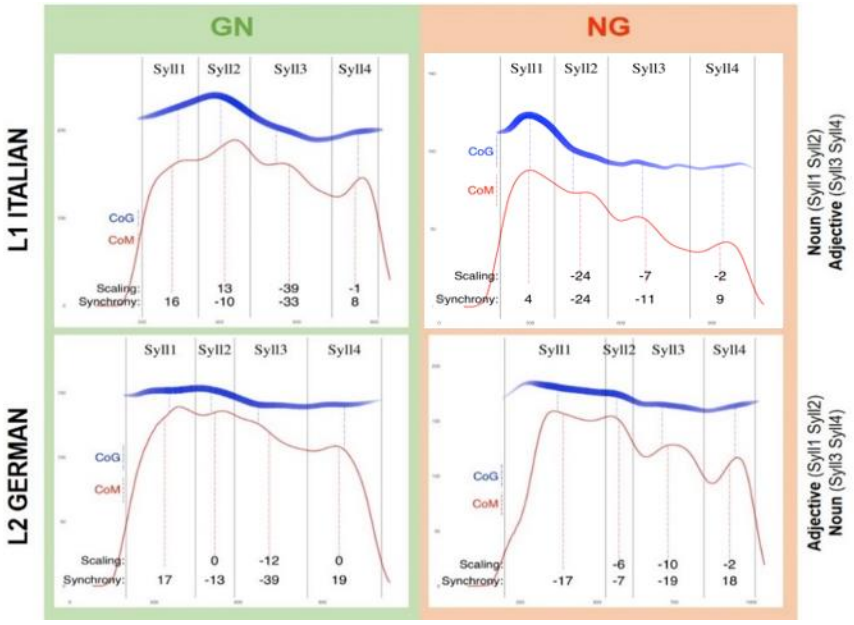


Figure 1. Periograms and periodic energy curves for noun phrases with different information structures in L1 Italian and L2 German. Gray solid lines signal syllable boundaries. Vertical dashed/dotted lines denote the centre of gravity (CoG, in blue) and the centre of mass (CoM, in red) between two syllable boundaries. Numerical values for each syllable indicate *scaling*, measured as the difference between F0 values at the CoM of consecutive syllables; and *synchrony*, measured as the distance between the centres (CoG minus CoM) within each syllable (negative values reflect falling F0, positive values rising F0).

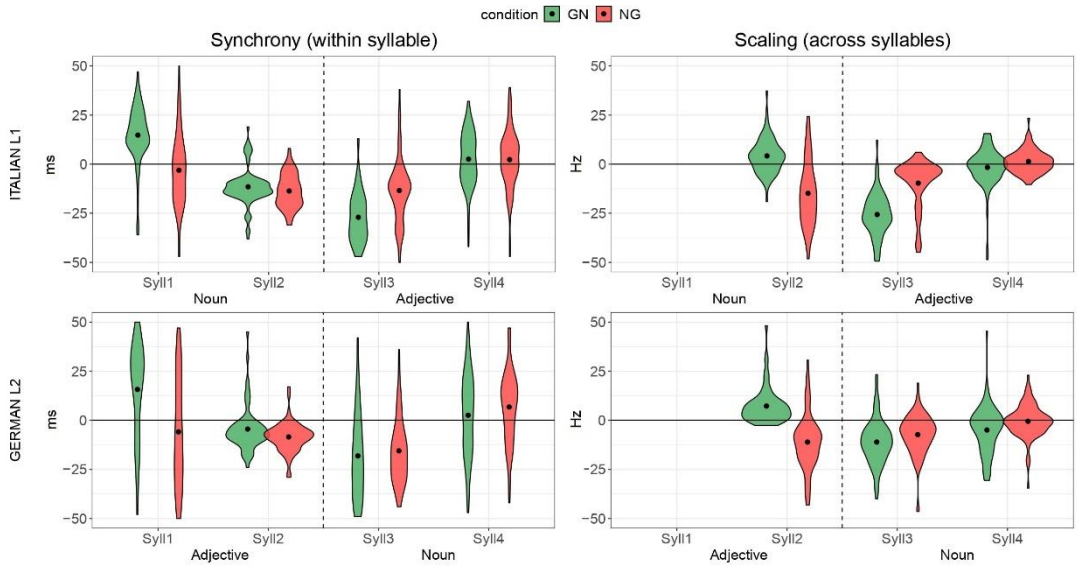


Figure 2. Violin plots for synchrony (left) and scaling (right) measures in noun phrases with different information structures averaged across all speakers in L1 Italian (above) and L2 German (below). Information structure is colour-coded, green violins refer to the given-new condition; red violins refer to the new-given condition. The x-axis displays the four syllables of the noun phrase. The y-axis shows values for synchrony (ms) and scaling (Hz). Values above the horizonal line crossing zero are positive, below it negative. Black points on the violins represent mean values.

Bibliography

AVESANI, C., BOCCI, G., VAYRA, M. & ZAPPOLI, A. (2015), Prosody and information status in Italian and German L2 intonation, in: *Il parlato in [italiano] L2: aspetti pragmatici e prosodici /[Italian] L2 Spoken Discourse: Pragmatic and Prosodic Aspect*, Milano, Franco Angeli, 93-116.

ALBERT, A., F. CANGEMI & M. GRICE (2018). Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. *Proceedings of the 9th Speech Prosody Conference*, June 2018, Poznan,1-5.

ALBERT, A., F. CANGEMI & M. GRICE (2020). *Periogram Projekt*. OSF. March 4. doi:10.17605/OSF.IO/28EA5.

BARNES, J. VEILLEUX, N., BRUGOS, A. & SHATTUCK-HUFNAGEL, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*. 3. 10.1515/lp-2012-0017.

CANGEMI, F., ALBERT, A. & GRICE, M. (2019). Modelling intonation: Beyond segments and tonal targets. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019, 572-576.

COUNCIL OF EUROPE (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Press Syndicate of the University of Cambridge.

LEVENE, H. (1960). Robust testes for equality of variances. In Olkin, I. (ed.) *Contributions to Probability and Statistics* (I. Olkin, ed.) Stanford Univ. Press, Palo Alto, 278–292.

SWERTS, M., KRAHMER, E. & AVESANI, C. (2002), Prosodic marking of information status in Dutch and Italian: a comparative analysis, *Journal of Phonetics*, 30, 4, 629-65.

Hesitations and Individual Variability in Italian Tourist Guides' Speech

Loredana Schettino[1], Simon Betz[2], Francesco Cutugno[3], Petra Wagner[2]
[1]University of Salerno, Italy; [2] Bielefeld University, Germany; [3]University of Naples "Federico II", Italy

In this study, we investigate hesitation strategies that tourist guides may use to manage speech with an emphasis on individual variability. It is by now recognized that speech disfluencies are not just occasional and idiosyncratic production errors, but naturally involved in the economy of speech and an integrative part of a language's grammatical organisation (Chafe 1980; Crocco and Savy 2003). Their use allows speakers to manage the online process of speech planning, coding, and articulation, correcting utterances in case of error and taking some extra time to organize the output message (Levelt 1989, Ginzburg et al. 2014). These "stalls" are marked by hesitations like silent pauses, lengthenings, and fillers which carry procedural communicative values, although lacking in propositional content. Their occurrence buys time for the speakers to manage their speech and for listeners to process information, conveying at the same time valuable information on speech planning, structuring, and speakers' disposition (Cataldo et al. 2019, Betz 2020). However, there is no evidence of speakers' deliberate control over their production (see Eklund 2004, Corley and Stewart 2008).

Previous investigations on Italian tourist guides' speech (Cataldo et al. 2019) have pointed out that such "disfluencies" may occur as a tool to structure discourse and gain visitors' attention, and that linguistic idiosyncratic behaviour may affect their production. The individual variability issue was also recently tackled in a study on German semi-spontaneous speech where speaker-specific hesitation strategies emerged (Betz and Lopez Gambino 2016). Moreover, McDougall and Duckworth (2017) highlighted the speaker-discriminating role of disfluency production, which provides a further tool for forensic phoneticians.

Given these findings, the proposed study delves deeper into the linguistic analysis of formal, phonetic, and functional aspects of hesitations occurring in a corpus of tourist guides' speech. We investigated whether different types of hesitations and their phonetic features correlate with different discourse functions and what individual strategies speakers may, more or less consciously, use when hesitating.

We performed a corpus-based analysis on a dataset from the C.H.R.O.M.E. corpus (Origlia et al. 2018). It consists of circa 80' semi-spontaneous speech by three female expert guides leading visits at San Martino's Charterhouse. Disfluency phenomena were annotated using a three-level annotation scheme: Disfluency Model, Disfluency Structure, Disfluency Function (see Shriberg 1994, Eklund 2004, Ginzburg et al. 2014). Phenomena falling into the set of "hesitation pauses" were associated with possible function/s according to their co-text. The types considered were: Silent Pauses (SP); Filled Pauses (FP, such as "ehm" or "eeh"); Lengthenings (LEN, as in "nell<aa> Certosa"); Lexicalized Filled Pauses (LFP, such as "let's say", "so"). Functions were classified as follows: Hesitative, referring to speech planning as hesitation basic function; Word searching, for items revealing difficulties' in retrieving target words; Structuring, when structuring discourse at syntax and information structure level; Focusing, when emphasizing upcoming semantically heavy elements (see Cataldo et al. 2019). The robustness of this categorization was tested measuring inter-rater reliability, Cohen's kappa was 0.732 ("substantial agreement", Landis and Koch 1977).

Results confirm and dig deeper into findings by Cataldo et al. (2019) reporting the emergence of idiosyncratic linguistic behaviours. The 1158 hesitation items occurring in the dataset are unevenly distributed across the three speakers (G01, G02, G03). Indeed, their G01's productions occur about twice as frequently as for the other two guides, see Table 1. Preliminary statistical computation confirms that speakers adopt different strategies in the choice of hesitation phenomena and their pragmatic functions. A Multinomial Logistic Regression Model was fitted in R, defining "type" as the dependent variable and "speaker" and "pragmatic function" as interacting independent variables. Then, a pairwise comparison among fixed levels as to the speaker effect on the hesitation choice showed the following significant results: compared to the other two speakers G01 uses more filled pauses and lengthenings, fewer lexical fillers; G02 fewer silent pauses; G03 fewer lengthenings ($p<.001$). Also, the interaction between hesitation type and function was found to be significant ($p<.001$), namely, as compared to the other guides, in G01's speech more lengthenings are used with focusing function and more filled pauses with structuring function, in G02's speech more lexical fillers are used to convey hesitative and word searching function, in G03's speech more silent pauses are used for hesitative function whereas fillers and lengthenings for word searching function.

To get general correlations between hesitations' formal and functional features besides inter-speakers' variability, mixed models considering "speaker" and "item" as random intercepts were employed – statements about cause-effect are avoided considering the number of variables at play in semi-spontaneous speech. Generalized Linear Mixed Models with hesitations' type as dependent variable and function as independent

variable showed that fillers are generally used with word searching function (SE: 0.19, $p < .001$) and not with focusing function (SE: 0.95, $p < .001$); lengthenings are also less used with focusing (SE: 0.38, $p < .001$), and structuring function (SE: 0.28, $p < .001$); conversely, lexical fillers are generally used to convey structuring (SE: 0.25, $p < .001$), and focusing function (SE: 0.28, $p < .001$). The correlation between hesitations' duration and the type was tested fitting a Linear Mixed Model, with duration as the dependent variable and type as the independent variable. It showed fillers (mean: 0.46 s) to be significantly longer than silent pauses (mean: 0.27 s; SE: 0.03, $p < .001$) and lengthenings (mean: 0.26 s; SE: 0.02, $p < .001$) To test the correlation between hesitations' duration and function, a Linear Mixed Model was built for each type defining the function as the independent variable. All hesitation types were found to be significantly longer when carrying out word searching function (***LEN*** means: 0.33 s vs. 0.20 s; SE: 0.02, p<.0001. ***FP*** means: 0.60 s vs. 0.30 s; SE: 0.04 p<.0001. ***SP*** means: 0.37 s vs. 0.27 s; SE: 0.04, p=.03). Interestingly enough, lengthenings were found to be significantly shorter when conveying focusing function (means: 0.16 s vs. 0.28 s; SE: 0.02, p<.0001).

To conclude, these results show how speakers use hesitation phenomena to manage their discourse and may choose different "hesitation strategies". Indeed, G01 prefers an 'on the fly' production, employing several filled pauses and lengthenings, also with, respectively, structuring and focusing function; G02 avoids silent pauses all together and tends to avoid lengthenings and filled pauses preferring lexicalized filled pauses for word searching and general hesitation functions as well as for structuring and focusing; whereas G03 adopts a more controlled, 'rhetorical' style, using mainly lexicalized filled pauses and silent pauses for emphasizing, structuring information, and for general hesitations, then, very few lengthenings and filled pauses occur when searching for a certain word. Overall, lengthenings and filled pauses were found to correlate with general hesitations and word retrieval problems, rather than for structuring and focusing functions, the latter being generally conveyed by lexicalized filled pauses instead. Noteworthy, the opposition between the word searching and focusing functions was found to be encoded in lengthenings' duration, respectively longer and shorter. Finally, although focusing on a restricted number of speakers, this study further uncovers the role of interspeaker variability when describing hesitations and the way they come into play in speech production and perception also in Italian, which may be a highly influential finding for a range of speech technological applications such as interactive speech synthesis,… Furthermore, it provides a better understanding of hesitation pauses' communicative functions and their contribution to discourse and communication.

Table 1 – Per minute, per word hesitation rate (first two columns) and instances of hesitations by Speaker

Speaker	n° hes/ minute	n° hes/ word	SP	FP	LEN	LFP
G01	20,12	0,16	78 (16%)	166 (34%)	144 (30%)	98 (20%)
G02	11,55	0,07	19 (6%)	42 (13%)	64 (19%)	210 (63%)
G03	12,10	0,08	53 (17%)	32 (10%)	32 (10%)	198 (63%)

References

Betz, S., Lopez Gambino, M. S. (2016). Are we all disfluent in our own special way and should dialogue systems also be?. Elektronische Sprachsignalverarbeitung (ESSV) 2016, 81.

Betz, S. (2020). Hesitations in Spoken Dialogue Systems. Ph.D. dissertation, Universität Bielefeld, Bielefeld.

Cataldo, V., Schettino, L., Savy, R., Poggi, I., Origlia, A., Ansani, A., Sessa, I., Chiera, A. (2019). Phonetic and functional features of pauses, and concurrent gestures, in tourist guides' speech. Studi AISV.

Chafe, W. (1980). Some reasons for hesitating. In Dechert, H.W., Raupach, M. (Eds), *Temporal variables in speech: Studies in Honour of Frieda Goldman-Eisler*. The Hague: Mouton, 169-180.

Corley, M., Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. Language and Linguistics Compass, 2(4), 589-602.

Crocco, C., Savy, R. (2003). Fenomeni di esitazione e dintorni: una rassegna bibliografica. In Crocco, C., Savy, R. & Cutugno, F. (Eds.), *API. Archivio di Parlato Italiano*, DVD.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. Ph.D. dissertation, Linköping University Electronic Press.

Ginzburg, J., Fernández, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9), 1-64.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 1, 159-74.

Levelt, W.J. (1989). *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT Press.

Lickley, R. J. (2015).Fluency and Disfluency.Redford, M.A.(ed.). *The handbook of speech production*,445-474.John Wiley & Sons.

McDougall, K. and Duckworth, M. (2017) Profiling fluency: an analysis of individual variation in disfluencies in adult males. *Speech Communication* 95: 16–27.

Origlia, A., Savy, R., Poggi, I., Cutugno, F., Alfano, I., D'Errico, F., Vincze, L. & Cataldo, V. (2018). An Audiovisual Corpus of Guided Tours in Cultural Sites: Data Collection Protocols in the CHROME Project. In *Proceedings of the 2018 AVI-CH Workshop on Advanced Visual Interfaces for Cultural Heritage*, Castiglione della Pescaia, Italy, 29 May 2018, 1-4.

Shriberg, E. (1994). Preliminaries to a theory of speech disfluencies. Ph.D. dissertation, University of California, Berkeley.

Forensic value of acoustic-phonetic features from Standard Dutch nasals and fricatives

Laura Smorenburg and Willemijn Heeren
Leiden University Centre for Linguistics

Although vowels generally outperform consonants in speaker discrimination, reports indicate that forensic voice analysts regularly use consonants in auditory-acoustic analysis [1]. However, research on the usefulness of acoustic-phonetic features from consonants in forensic speaker comparisons (FSC) is scarce. We investigated the forensic value of consonants that are highly frequent in Dutch and are therefore likely to be available in forensic material [2]: fricatives (/s x/) and nasals (/n m/). Fricatives are characterised by frication noise at higher or mid-range frequencies, depending on the place of articulation, whereas nasals are characterised by low-frequency energy due to nasal damping. Reports show that place of articulation and uvular trill in the velar/uvular fricative /x/ is strongly associated with region [3] and that sibilant fricative /s/ can carry speaker information such as gender, class, and sexual orientation [e.g. 4, 5]. Subsequent research has shown that /s/ is indeed speaker-specific in Dutch, meaning it has low within and high between-speaker variability [6]. Similarly, nasal consonants exhibit high speaker-specificity because of the nature of a nasal; the involvement of the relatively rigid nasal cavity, which has different shapes and sizes between speakers, results in high between-speaker but low within-speaker variation for nasals [7, p.135]. Because acoustic-phonetic analysis is prevalent in FSC [8], we investigated the forensic value of acoustic-phonetic features from Dutch nasals and fricatives in conversational telephone speech using the statistical framework used in FSC. Based on earlier work on Dutch (nonsense) read speech [6], we hypothesized that /n/ will outperform /m/ and that nasals outperform fricatives in speaker discrimination.

Method

Materials and acoustic analysis. Landline telephone conversations (bandwidth 340-3400 Hz) from adult male speakers of Standard Dutch were analysed [Spoken Dutch Corpus: 9]. From the same 62 speakers, we annotated 3,561 /s/ tokens (per speaker: M = 57, SD = 24), 3,836 /x/ tokens (per speaker: M = 62, SD = 31), 4,676 /n/ tokens (per speaker: M = 74, SD = 28), and 3,654 /m/ tokens (per speaker: M = 58, SD = 24). For fricatives, the following features were extracted per token: duration (log10-transformed), centre of gravity (CoG), standard deviation (SD), skewness (SKW), kurtosis (KUR), and spectral tilt. CoG was also measured in five non-overlapping windows of 20% of a token's duration, after which a cubic polynomial fit was made to capture the dynamics of CoG, resulting in four coefficients. For nasals, we also measured the second and third nasal formants (N2, N3), and their bandwidths (BW2, BW3). N2 and N3 were also captured dynamically, in the same way as CoG.

Statistical analysis. Speaker discriminability was established with likelihood ratios (LR), which reflect the ratio of the probability of the evidence under the hypothesis that two speech samples come from the same speaker (SS) to the probability of the evidence under the hypothesis that two speech samples come from different speakers (DS). The analysis was performed using a MATLAB implementation [10] based on the LR algorithm proposed in [11], where within-speaker variation is modelled as a normal distribution and between-speaker variation is modelled with a multivariate kernel density. LR systems were built for each consonant, using acoustic-phonetic features as parameters. Highly correlating features may inflate the strength of evidence, so a maximum correlation was set at $r = .50$. For /s/ and /x/, this resulted in the following parameters: duration, CoG, SD, Kur, and the three dynamic CoG coefficients. For /n/ and /m/, we used the same parameters for a direct comparison with the fricatives and included the nasal formants and bandwidths in a separate system.

Per system, the 62 speakers were divided into a development (N=22), reference (N=20), and test set (N=20). First, SS and DS LRs were computed for the development set. Not all speakers had multiple recordings, so the tokens per speaker were divided in half to generate SS

comparisons. For the development set, this resulted in 22 SS and 231 DS comparisons. The LR scores from these comparisons were used to obtain calibration parameters (shift, slope) for the test set. LLRs were then obtained and calibrated for the test set. To reduce sampling effects, 10 iterations were used in which the development, reference, and test sets were sampled at random. The systems' performance was assessed through SS and DS LLRs and the log-likelihood-ratio costs (C_{llr}), which reflects the degree of accuracy of the system's calibrated decisions. Median LLRs and C_{llr}s over iterations were obtained using R package *sretools* [12].

Results

Table I displays the results. An LLR of 1 means that the evidence is 10 times more likely under the same-speaker (SS) hypothesis and an LLR of –1 means it is 10 times more likely under the different-speaker (DS) hypothesis. E.g., the LLR_{SS} of 1.52 means that the evidence is 33 times more likely under the SS hypothesis than the DS hypothesis. For C_{llr}, closer to 0 is better.

Table I. Median SS and DS LLRs and C_{llr}s

	Static parameters			Dynamic parameters			Static nasal-specific parameters			Dynamic nasal-specific parameters		
	LLR_{SS}	LLR_{DS}	C_{llr}	LLR_{SS}	LLR_{DS}	C_{llr}	LLR_{SS}	LLR_{DS}	C_{llr}	LLR_{SS}	LLR_{DS}	C_{llr}
/s/	1.52	–2.36	0.52	0.25	–0.10	0.91						
/x/	0.74	–0.20	0.82	0.26	–0.03	0.96						
/n/	0.74	–0.60	0.67	0.43	–0.08	0.87	1.55	–1.54	0.55	0.13	–0.08	0.96
/m/	0.85	–0.50	0.71	0.21	–0.07	0.93	1.05	–0.78	0.70	0.03	0.01	0.99

Discussion and conclusion

Results indicate that /s x n m/ have forensic value, but that the extracted acoustic-phonetic features differ in their discriminatory power. Static acoustic-phonetic features contained more speaker information than dynamic acoustic-phonetic features. This is perhaps due to contextual influences in these short consonants leaving little speaker-specific information in the dynamics. Nasals performed better with static nasal-specific features. Against expectations, we found that /s/ outperformed the other consonants, even though it was sampled from telephone speech and its spectral peak falls outside of the telephone band.

Acknowledgement NWO VIDI grant (276-75-010) supported this work.

References

[1] Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law, 18*(2), 293–307.

[2] Luyckx, K., Kloots, H., Coussé, E., & Gillis, S. (2007). Klankfrequenties in het Nederlands. In *Tussen taal, spelling en onderwijs* (pp. 141–154). Academia Press.

[3] Harst, S. Van der, Velde, H. Van de, & Schouten, B. (2007). Acoustic characteristics of Standard Dutch /x/. *Proceedings of the 16th ICPhS*, 1469–1472.

[4] Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *J.Phon., 34*, 202–240.

[5] Stuart-Smith, J. (2007). Empirical evidence for gendered speech production: /s/ in Glaswegian. *Change in Phonology: Papers in Laboratory Phonology, 9*, 65–86.

[6] Van den Heuvel, H. (1996). *Speaker variability in acoustic properties of Dutch phoneme realisations*, Radboud Universiteit, Nijmegen.

[7] Rose, P. (2002). Forensic Speaker Identification. In *Sciences New York* (Vol. 20025246).

[8] Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: Second survey. *International Journal of Speech, Language and the Law*, *26*(1), 1–20.

[9] Oostdijk, N. H. J. (2000). Corpus Gesproken Nederlands. *Nederlandse Taalkunde, 5*, 280–284.

[10] Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation. [software].

[11] Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *J. of the Royal Stat. Soc. Series C: Applied Statistics*, *53*(1), 109–122.

[12] Van Leeuwen, D. (2011). SREtools: Compute performance measures for speaker recognition.

System performance and speaker individuality in LR-based forensic voice comparison

Bruce X. Wang, Vincent Hughes and Paul Foulkes
{xw961, vincent.hughes, paul.foulkes}@york.ac.uk
Department of language and linguistic science, University of York

1. Introduction

The speaker-discriminatory power of segmental phonetic features has been explored in numerous studies within the likelihood-ratio (LR) framework (Hughes et al., 2016; Morrison, 2009; Rose & Wang, 2016; Zhang et al., 2011). Many studies use vowel formants to test overall system validity. In general, combining more segmental phonetic features (e.g. multiple vowels) or parameters (e.g. multiple formants) improves system validity. However, few studies have explored system stability using different phonetic parameters, and how individual speakers are affected when different systems are used (but see Morrison et al. 2011; Wang et al., 2019). The current study addresses these issues by considering the performance of systems and the effects on individual speakers (i.e. validity and stability).

2. Methods

Filled pauses (FPs) (*um*) from 90 SSBE speakers (DyViS corpus, Nolan et al., 2009) were used. Quadratic polynomials were fitted to the F0, F1, F2 and F3 trajectories of the vowel portion of *um.* The coefficients and both vocalic and nasal durations were used for LR comparisons. Five systems were tested: F1, F2, F3, vocalic/nasal durations, and the combination of all four. In each system, 25 speakers were randomly sampled into the test, training and reference sets. The multivariate kernel density formula (MVKD, Aitken & Lucy, 2004) was used to compare the same-speaker (SS) and different-speaker (DS) pairs of test and training data to produce a series of test and training scores. The training scores were then used to build a logistic regression model (Brümmer et al., 2007; Morrison, 2011) that was applied to test scores to produce calibrated Log_{10} LRs (LLR). The experiment was replicated 100 times, sampling training and reference speakers. This provides insight into system stability, because training and reference speakers were treated as the system, and test speakers were used as suspect and offender samples. Using the same test speakers across 100 replications enables us to explore the effect of different systems (i.e. different configurations of training and reference speakers) on individual speakers. System performance was evaluated using the log LR cost function (C_{llr}, Brümmer & du Preez, 2006), while results for individual speakers were assessed using mean LLRs with a modified zoo plot (Doddington et al., 1998) and root-mean-square error (RMSE). Animal groups, i.e. *chameleons, phantoms, doves,* and *worms*, adapted from Dunstone and Yager (2009) were used in the zoo plot. Instead of upper and lower quartiles of scores that are traditionally used in ASR, the zoo plot thresholds were adjusted based on the LLR verbal expression (Champod and Evett, 2000; Table 1). This is because calibrated LLRs were used for zoo plots, and LLRs are comparable between speakers and across systems, which are different from comparison scores used in ASR systems.

Animal group	SS LLR	DS LLR
Phantoms	≤ 0	≤ -1
Worms	≤ 0	≥ -1
Doves	≥ 1	≤ -1
Chameleons	≥ 1	≥ 0

Table 1. LLR threshold for animal groups

3. Results

Figure 1 shows the C_{llr} range across 100 replications in five systems. Combining all parameters produces the best validity, with the lowest C_{llr} across replications of 0.14. This is followed by F2 (min. C_{llr} = 0.37), F3 (min. C_{llr} = 0.60), F0 (min. C_{llr} = 0.65), F1 (min. C_{llr} = 0.69) and duration (min. C_{llr} = 0.72). The F1 system yields the largest overall range (OR = 0.47), followed by the combined system (C_{llr} OR = 0.17), F3 (C_{llr} OR = 0.12), Duration (C_{llr} OR = 0.13) and F2 (C_{llr} OR = 0.04). Figure 2 reveals a lack of *worm* and *chameleon* groups across all systems. Speakers in general yield better performance with more parameters, as the majority of speakers (18 out of 25) shift to the *dove* group (right top) in the combined system. Speakers' performance varies across single parameter systems; speakers 72, 54 and 114 fall in the *phantom* group in the Duration and F1 systems, while they shift to the *dove* group in the combined

system. However, speaker 120 is classified in the *phantom* group in the F2 system but does not shift to the *dove* group in the combined system. Some other speakers, e.g. 48 and 53, can be well-separated using F2 alone, but yield misleading LLRs in the F1 and F3 systems. Moreover, the patterns of speakers 48 and 53 show that using combined parameters does not necessarily contribute to the magnitude of the strength of evidence. For SS comparisons, all speakers tend to fluctuate most in the combined system and least in single parameter systems. Meanwhile, the majority of speakers show little fluctuation (RMSE values vary between ca. 0.1-0.2) across single parameter systems (e.g. speakers 8, 13, 20). For DS comparisons, speakers fluctuate much more in their DS LLRs across different systems. This is due to the fact that between-speaker variance is likely to be larger than within-speaker variance (Rose & Morrison, 2009). Only speakers 36 and 90 seem comparatively stable across different systems, where the DS RMSE values vary between 0 and 2.5. In DS comparisons, the 19 speakers tend to fluctuate most in F2 systems (e.g. speakers 13, 20, 21), while the other six - 8, 46, 48, 51, 53 and 77 - yield the most variable performance in the combined system.

4. Conclusion

The current study used replications to explore system performance and individual speakers' behaviour, which provides novel insights for forensic speech science. The experiments show that combining multiple parameters contributes to overall system performance, and the performance of individual speakers is system-specific, i.e. training/reference speakers and parameters used.

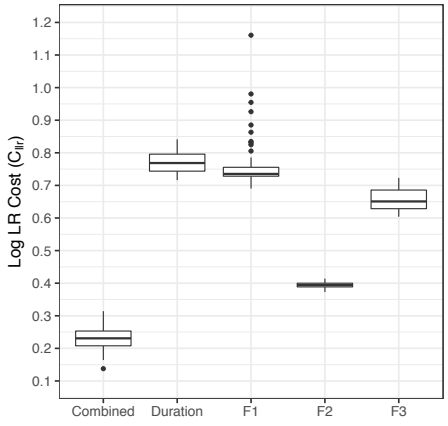


Figure 1. C_{llr}s in six systems

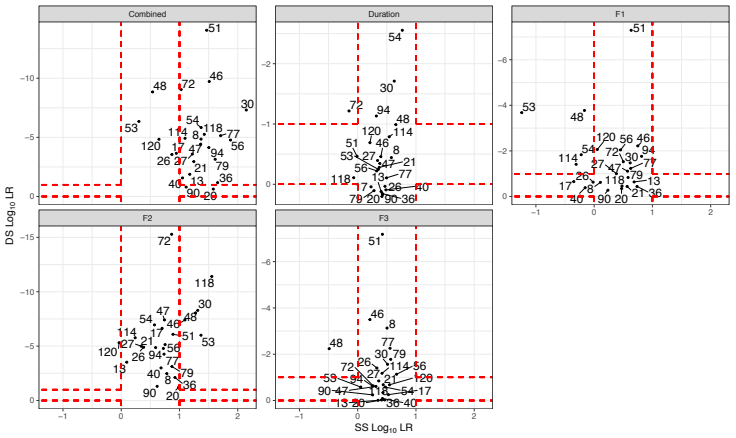


Figure 2. Zoo plot of 25 test speakers in six systems.

References

Aitken, C. G. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(1), 109–122. https://doi.org/10.1046/j.0035-9254.2003.05271.x

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2072–2084. https://doi.org/10.1109/TASL.2007.902870

Brümmer, N., & du Preez, J. (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, *20*(2–3), 230–275. https://doi.org/10.1016/j.csl.2005.08.001

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. *Proc. Int'l Conf. Spoken Language Processing*, 4.

Dunstone, T., & Yager, N. (2009). *Biometric system and data analysis: Design, evaluation, and data mining*. Springer.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech Language and the Law*, *23*(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

Morrison, G. S. (2009). Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aI/. *International Journal of Speech Language & the Law*, *15*(2), 249–266. https://doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). *Speech Communication*, *53*(2), 242–256. https://doi.org/10.1016/j.specom.2010.09.005

Morrison, G. S., Zhang, C., & Rose, P. (2011). An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic Science International*, *208*(1–3), 59–65. https://doi.org/10.1016/j.forsciint.2010.11.001

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language & the Law*, *16*(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31

Rose, P., & Morrison, G. S. (2009). A response to the UK Position Statement on forensic speaker comparison. *International Journal of Speech Language & the Law*, *16*(1), 139–163. https://doi.org/10.1558/ijsll.v16i1.139

Rose, P., & Wang, B. X. (2016). *Cantonese forensic voice comparison with higher-level features: Likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone*. 326–333. https://doi.org/10.21437/Odyssey.2016-47

Wang, X. B., Hughes, V., & Foulkes, P. (2019). The effect of speaker sampling in likelihood ratio based forensic voice comparison. *International Journal of Speech Language and the Law*, *26*(1), 97–120. https://doi.org/10.1558/ijsll.38046

Zhang, C., Morrison, G. S., & Thiruvaran, T. (2011). Forensic voice comparison using Chinese /iau/. *Hong Kong*, ICPhS (pp. 2280-2283).

Poster Presentations

Italian monozygotic twins' speech: a preliminary forensic investigation

Alice Albanesi*, Sonia Cenceschi**, Chiara Meluzzi***, Alessandro Trivillini**

*Institute of Forensic Sciences, Milano, Italy

** Digital Forensic Service, DTI, University of Applied Sciences and Arts of Southern Switzerland

*** University of Pavia, Pavia, Italy

The main purpose of this preliminary study is to investigate whether it is possible to distinguish a speaker from his co-twin in compressed audio recordings. The twins speech similarity degree depends on the changing sum of an anatomical inheritance with environmental and social factors which contribute to sculpt their personality [1]. However, despite the debate on these factors, the monozygotic twins' voices represent a crucial point in forensics: beyond the practical cases involving siblings, they represent the most extreme physical similarity between two different speakers, and consequently the very lowest limits of between-speaker variation [2-3]. As a consequence, they are an excellent starting point to study a number of key topics in forensics such as the acoustic features influencing speaker's recognition accuracy, or the auditory discrimination of voices.

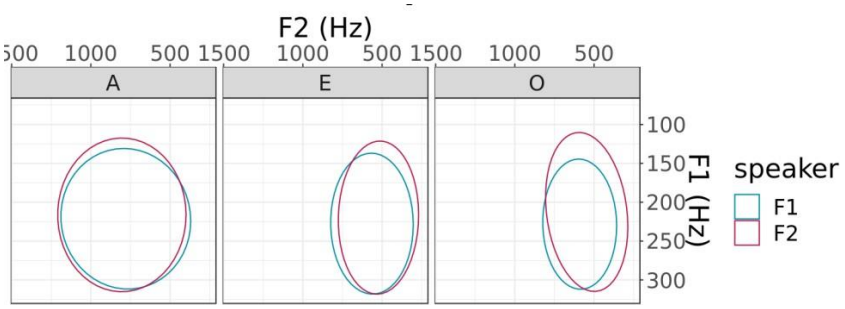
A rich international scientific literature exists [e.g. 4—8] but it tends to drastically decrease for Romance languages, where *more forensic corpora should be created, for example considering low quality recordings* [3]. Despite the presence of several Italian speech corpora [e.g. 9—11], nothing available seems to exist concerning twins' speech. A few investigations focus on the Italian twins perceptual acoustic discrimination [e.g. 12-13] but their results are still limited due to the low number of samples and because the results do not deal with the forensic perspective.

This first explorative work aims at investigating the speech of a small selected corpus of twins' speech. The first question to be addressed is whether and according to which variables Italian twins speech could be distinguished. Secondly, whether and to what extent these (dis)similarities are reduced or enhanced in controlled vs. spontaneous speech.

This preliminary study is conducted on a small corpus consisting in 6 pairs of monozygotic twins (3M and 3F), with the same sociolinguistic characteristics (high level of education, born and living in the north-west of Italy, L1 Italian) aged between 20 and 25 y.o. Each speaker was asked to record a list of 31 elicited sentences [14] and undergo a short interview. The similarity of prosodic profile was aimed at reducing the involved variables as much as possible to ensure repeatability, and to lay the ground for future studies.

In forensic contexts, audio files are often collected with different technologies, and are typically characterized by very low quality because of difficult and noisy recording conditions (e.g. recording with micro recorder in a restaurant), narrow frequency bands, and compressed formats (e.g. telephone interception). In order to mimic a forensic setting, we recorded with four different smartphone, creating m4a files (48-44 kHz, 16 bit), later converted to *.wav 44.1 kHz 16 bit for Praat processing (spectral band cut at 15kHz). Recordings have been realized in a silent room, but without specific details regarding environmental soundproofing. Vowels /a/, /e/ and /o/ were manually annotated on a PRAAT tier (both in sentences and interviews). In order to emulate as closely as possible forensic conditions (short or partially unusable audio), all possible vowels were taken [15] in order to start observing their overall variability on each task. For setting the vowel's left and right boundaries we based on the beginning and end of the second formant (not selecting if it wasn't clear to listening) obtaining about 440 vowels for typology on each task. After the annotation, we automatically extracted the following acoustic parameters at the midpoint using a Praat script: F0, F1, F2 and F3 of the target vowels. Then, formants values have been visually inspected through the web application Visible Vowels [16] as suggested in [17], with the statistical analysis carried out with IBM SPSS 20.

A first analysis shows that all pairs present almost identical distributions for all vowels with slight differences in spontaneous speech. F0 and F3 are really similar, with minimal oscillation that can be explained by expressive variations. The figure shows an example of strong similarities in the F1-F2 unnormalized distribution for the first female pair in controlled speech. In spontaneous speech the distributions remain displaced in the same way, with a very slight shift of the lower F2 boundary towards the low frequencies for /e/ and /o/.



In the following months we aim at expanding the corpus and integrating our analysis with other acoustic parameters such as jitter and shimmer, but also dynamic variation of F2, since the second formant has proved to be a robust indicator of between-speaker variation in sociophonetics as well as in clinical phonetic research [18]. Moreover, vowels observation will be refined in order to observe in detail the proximity to certain consonants and their dynamic behavior (over 7 time steps). A parallel investigation concerns the perceptual aspect: once the analysis of the corpus has been completed, a perceptual test will be prepared to understand whether the hearing system validates (or not) the results.

References

[1] Nolan, F., & Oh, T. (1996). Identical twins, different voices. International Journal of Speech, Language and the Law, 3(1), 39-49.

[2] Loakes, D. (2008). A forensic phonetic investigation into the speech patterns of identical and non-identical twins. International Journal of Speech, Language and the Law, 15(1), 97-100.

[3] Fernández, E. S. S. (2013). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. Procedia-Social and Behavioral Sciences, 95, 59-67.

[4] van Braak, P., & Heeren, W. F. L. (2015). "Who's calling, please?" Is there speaker-specific information in twins' vowels?, Bachelor's thesis.

[5] Johnson, K., & Azara, M. (2000). The perception of personal identity in speech: Evidence from the perception of twins' speech. Unpublished manuscript.

[6] Sebastian, S. (2013). An investigation into the voice of identical twins. Otolaryngology online journal, 3(2), 9-15.

[7] Zuo, D., & Mok, P. P. K. (2015). Formant dynamics of bilingual identical twins. Journal of Phonetics, 52, 1-12.

[8] Van, W. G., Vercammen, J., & Debruyne, F. (2001). Voice similarity in identical twins. Acta oto-rhino-laryngologica Belgica, 55(1), 49-55. PERCEZIONE

[9] Falcone, M., & Barone, A. (2000). FOCUS: un corpus vocale di voci simili per lo studio della identificazione del parlatore in ambito forense. 28 Convegno AIA.

[10] Falcone, M., & Gallo, A. (1996, October). The" siva" speech database for speaker verification: Description and evaluation. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3, pp. 1902-1905). IEEE.

[11] Cresti, E., & Moneglia, M. (Eds.). (2005). C-ORAL-ROM: integrated reference corpora for spoken romance languages (Vol. 15). John Benjamins Publishing.

[12] Giannini, A. (1989). Test acustico-percettivo su voci di gemelli. In atti del xvii convegno nazionale aia (pp. 427-432).

[13] Gedda, L., Bianchi, A., & Bianchi-Neroni, L. (1958). La Voce dei Gemelli—I. Prova di identificazione intrageminale della voce in 104 coppie (58 MZ e 46 DZ). Acta geneticae medicae et gemellologiae: twin research, 4(2), 121-130.

[14] Prieto, P., Borràs-Comes, J., & Roseano, P. (Coords.) (2010-2014). Interactive Atlas of Romance Intonation. Web page: <http://prosodia.upf.edu/iari/>.

[15] Rhodes, R. W. (2012). Assessing the strength of non-contemporaneous forensic speech evidence (Doctoral dissertation, University of York) p.121.

[16] Heeringa, W., van de Velde, H. (2018). Visible Vowels: a Tool for the Visualization of Vowel Variation. In Proceedings CLARIN Annual Conference 2018, 8 - 10 October, Pisa, Italy.

[17] Cenceschi, S., Meluzzi, C. (2020). The variability of vowels' formants in forensic speech. In IEEE Instrumentation & Measurement Magazine, Vol.24 (in press).

[18] Meluzzi, C. (2021, to appear). Sound Spectrography, in M.J. Ball (ed.) Manual of Clinical Phonetics, London: Routledge, pp. 418-443.

A 3D model of linguopalatal contact for VR biofeedback.

Chiara Bertini*, Paola Nicoli*, Niccolò Albertini*, Chiara Celata°
* Scuola Normale Superiore, Pisa
° Università degli studi di Urbino 'Carlo Bo'

In this paper we describe a 3D model of the linguopalatal contact issued from real multilevel data for the simulation in a virtual reality (VR) environment of the mechanisms underlying the production of lingual sounds. The outcome is an animation that can be experienced within a Unity 3D graphics engine, desktop or immersive environment.

Motivation: The model was developed within a project on speech motor disorders aimed at developing rehabilitation techniques based on VR visual biofeedback (Barone 2017). Modelling the spatiotemporal dynamics of linguopalatal contact is important in the context of many speech pathologies for both diagnosis and rehabilitation. Several biomechanical models of tongue movements exist that are based on mathematical models of muscular actions and their interactions (Moschos et al. 2011, Lloyd et al. 2012, Wrench & Balch 2015 among others). The model developed in this project takes an opposite kinematic perspective: the goal is to produce a patient-specific model starting from real articulatory data issued from specific experimental settings. As a matter of fact, our model exploits the information about the positioning of the tongue with respect to the palate, and this information is obtained from real data acquired by means of a digital ultrasound device for tongue imaging (UTI) and an electropalatograph (EPG) in synchronized combination (Spreafico et al. 2015, Celata et al. 2018). Furthermore, we do not model the movement of the tongue in general but, more specifically, the contact between the active (tongue) and passive (palate) articulator in the production of lingual sounds.

Data acquired: The 3D reconstruction of the palate has been obtained by acquiring and processing midsagittal and transversal ultrasound images of the speaker's palate; subsequently, the plaster cast of the palate and the artificial palate have been scanned. By superimposing the two spatial information, the 3D anatomy of the speaker's upper oral cavity has been reconstructed. An echogenic object of known size and shape (biteplane) has been used as reference for both the alignment of the virtual structures of the oral cavity and the analysis of the multilevel data obtained from different experimental sessions.
UTI data consisted in the discrete sampling over time of the mid-sagittal profile of the tongue during speech production. The Micro Speech Research Ultrasound system developed by Articulate Instruments Ltd was used, equipped with a micro-convex transducer (10mm; 5-8MHz; max FOV 150°); the software for the analysis was Articulate Assistant Advanced (AAA). At each ultrasound frame, a maximum of 42 discrete points, corresponding to the lines of sight of the ultrasound probe, are used to reconstruct the tongue midsagittal upper contour. The reading of the tongue profile data is therefore n <= 42 positions supplied as coordinates (in mm) in the x-y plane at each ultrasound frame.
EPG data consisted of binary information about presence/absence of contact (value 1 or 0) between the tongue and the 62 sensors arranged on the artificial palate worn by the speaker. This information is acquired by the WinEPG system by Articulate Instruments Ltd and analysed through the same software AAA.
Both UTI and EPG data were sampled at a frequency of 100 Hz. The data were then arranged in tabular form so as to have, for each row, the reference time point and the sequence of positions (for the UTI data) and contacts (for the EPG data).

Rigging and skinning of the model: The tongue model has been created by using chains of bones (i.e., chains of movement units for a 3D object during an animation), whose sum represents the skeleton of the virtual object (the tongue).
The skeleton was made up of 9 chains, each of which consisted of 42 bones. Each bone was positioned in the virtual space according to the spatial coordinates recorded by the articulatory instruments, and was directed towards the bone immediately in front of it. The central chain is the one responsible for receiving the positional information coming from the UTI data. When the acquired UTI data were < 42 (e.g. when the tongue is retracted and does not intersect the front rays), the missing data were filled through the Bézier interpolation algorithm. The central chain was therefore animated at each frame by the acquired UTI positional data and then transmitted this positional information to the side chains. The side chains corresponded each to a different sensor line of the artificial palate. When a sensor was contacted at a given time, the closest bone of the corresponding chain was detected and the position of the sensor was attributed to that bone. Since the same sensor was not always contacted by the same

area of the tongue, a function was created that evaluate which part of the tongue surface was most likely to contact a given sensor based on proximity.
The skeleton was first tested on a very simple polygonal skeleton (mesh), and subsequently incorporated into that of the definitive virtual model of the tongue. The skinning process was the definition of which vertices were influenced by which bone and with what weight. The automatic association provided by the software was manually corrected and validated.

Scripting and animation: The software used to virtually create the oral cavity and animate the tongue was 3ds Max 2019. A script was implemented to automate the process of acquiring data from files and creating the animation. The script is executable within the 3ds Max 2019 program via .mcr file so that it can be called up directly from the user interface.

Management of the mandibular movement: To get an estimate of the angle of mandibular rotation, light sensors were positioned at specific points of the speaker's face. Using the Intel® RealSense ™ D400 camera for facial recognition, the movements of the sensors during the production of the speech stimuli were recorded. The resolution of the camera allowed a good estimate of the angle of rotation of the jaw. These measures were inserted as correction values directly in the virtual display interface.

Visualization in a virtual environment: The digital elements necessary for the creation of the interface were the tongue, the palate and the jaw. The palate was derived from the acquisition of a real cast by laser scanning and the production of a model for Unity 3D. For the mandible model, the mesh available in Artisynth (Lloyd et al. 2012) was used. Once extracted, the polygonal object underwent a modeling process on 3DS Max. For the tongue, the Artisynth model was initially used but subsequently a new mesh was modelled according to the needs of the rigging calibration with the data obtained through the articulatory instruments. After defining the geometric structure, the display was optimized according to standard procedures.
The last phase focused on the development of the application using Unity 3D in order to obtain an immersive 3D visualization allowing interactivity and customization of parameters and animations (see figure below).

Corpus: The speech corpus used for the development of the prototype is very small and will have to be enlarged. The testing and validation were carried out on a set of 12 bisyllabic pseudo-words produced by a female speaker. These included all Italian vowels and, for the consonants, alveolar and velar stops, the lateral approximant and the alveolar trill.

Future perspectives: We will discuss the multiple possibilities for further development of the current prototype, from the introduction of an automatic system based on neural networks for the correction of the experimental noise, to gamification options to facilitate the use of the app by children. We will also show possible uses in speech therapy and speech teaching.

Barone V. (2017-2020). *Disturbi motori nel parlato e biofeedback visivo: Simulare i movimenti articolatori in 3D*. Progetto finanziato da Fondazione Pisa presso Scuola Normale Superiore di Pisa.

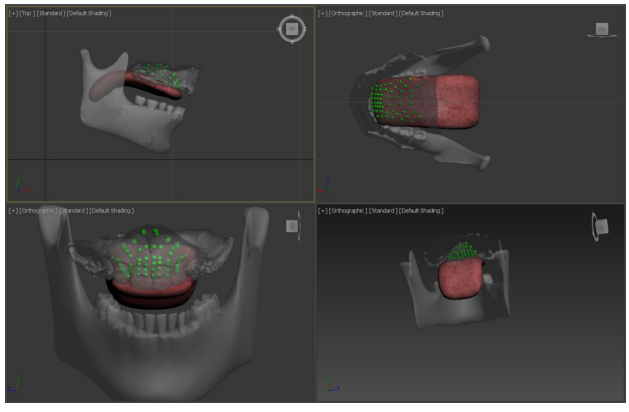
Celata / Vietti / Spreafico (2018). An articulatory account of rhotic variation in Tuscan Italian: synchronized UTI and EPG data. In *Romance Phonetics & Phonology*, Oxford University Press.

Lloyd / Stavness / Fels (2012). ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery*, 355-394.

Moschos / Nikolaidis/ Pitas / Lyroudia (2011). A virtual anatomical 3D head, oral cavity and teeth model for dental and medical applications. In *Man-Machine Interactions* 2, 197-206.

Spreafico / Celata / Vietti / Bertini / Ricci (2015). An EPG + UTI study of Italian /r/. In *Proceedings of 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, 2015.

Wrench / Balch (2015) Towards a 3D Tongue model for parameterising ultrasound data. In *Proceedings of 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, 2015.



An instant of the animation.

Sull’insegnamento della pronuncia italiana negli anni sessanta a bambini e a stranieri

Silvia Calamai (Università degli Studi di Siena)
Cecilia Valentini (Università degli Studi di Siena)

Nel corso della storia dell’italiano la questione della pronuncia è stata lungamente dibattuta, non senza polemiche e prese di posizione talvolta intrise di pregiudizi e campanilismo. Gli studiosi che se ne sono occupati possono essere raggruppati in due filoni: uno, etichettabile come purista, prescrive una pronuncia basata sull’italiano di Firenze (e, in alcuni periodi storici, anche di Roma); l’altro ha un approccio più tollerante ed è orientato soprattutto alla descrizione delle varie pronunce locali. Un modello di standard parlato sarebbe il cosiddetto fiorentino emendato, basato sulla pronuncia colta di Firenze, privata però di alcuni tratti marcatamente locali (ad esempio la cosiddetta gorgia toscana). Tale modello costituisce un punto di riferimento normativo ed è consigliato dai manuali di ortoepia (quali Malagoli 1905 e Camilli & Fiorelli 1965), dai corsi di dizione (Fiorelli 1964; Tagliavini 1965), dai trattati di fonetica e dai vocabolari (Migliorini, Tagliavini & Fiorelli 1969); tuttavia, dato il suo carattere artificioso, non ha avuto applicazione nell’insegnamento scolastico e in pratica non è appreso da nessun parlante come lingua materna. A questo standard si uniformano i “professionisti della parola” (attori teatrali, doppiatori, annunciatori radiofonici e televisivi), anche se al giorno d’oggi pronunce alternative, specialmente romane o settentrionali, sono molto più tollerate (Galli de’ Paratesi 1984: 53; Canepari 1999^{2}: 21).

Da questa prospettiva, ispezionare le indicazioni dei manuali e dei corsi di ortoepia significa anche compiere un cammino nella storia culturale e linguistica italiana. All’interno del progetto *Ti racconto in italiano*, promosso dall’Istituto Centrale per i Beni Sonori ed Audiovisivi in collaborazione con l’Università di Siena, sono stati digitalizzati e sono ora in corso di indicizzazione due corsi di pronuncia: *Corso di ortoepia*, a cura del Centro Nazionale Sussidi Audiovisivi, con testo di Antonio Mura e dizione di Giovan Battista Arista (Roma, Editrice italiana audiovisivi, 1960); *La pronunzia dell’italiano insegnata agli stranieri*, a cura di Umberto Pittola (Roma, Editrice italiana audiovisivi, 1961). L’intervento mira a contestualizzare questi due documenti proprio nell’ambito della storia della pronuncia dell’italiano e dedica particolare attenzione al trattamento dei fenomeni fonetici che rappresentano (o rappresentavano) punti di crisi nell’insegnamento della pronuncia “corretta”. Il materiale documentario offre infatti la possibilità di riflettere sulla questione della norma, in particolare sulla varietà di pronuncia proposta come modello da imitare. Le particolari condizioni di diffusione dell’italiano, che in molte aree e per la maggior parte delle persone restò una lingua esclusivamente scritta per lungo tempo, sono la causa della pesante influenza della grafia sulla pronuncia; a partire dall’unità d’Italia, inoltre, il modello toscano, pur persistendo a livello di norma, entra in forte concorrenza con la pronuncia propria dei grandi centri politici ed economici italiani (Roma, Milano, Torino), che nella seconda metà del Novecento conosce poi una larga diffusione ad opera dei mezzi di comunicazione di massa.

Tali caratteristiche appaiono chiaramente nell’analisi dei due corsi di pronuncia. Nel primo la voce del maestro enuncia le regole e offre numerosi esempi di dizione, anche servendosi di brani poetici, facendoli ripetere per esercizio a un gruppo di bambini. Gli allievi scelti non hanno tutti una pronuncia perfetta, per cui vengono di volta in volta corretti; questo fatto, come viene precisato nell’*Avvertenza* al testo che accompagna la pubblicazione, è pensato come uno stimolo all’insegnante che si serve di tale corso per trovare “difetti” (sic) simili da correggere nei propri scolari. La pubblicazione è chiaramente destinata ai maestri come sussidio durante le lezioni, come

indicano l'uso di una terminologia propria della scuola elementare (i nomi delle lettere *bi*, *ci*, *elle*, *emme*, *enne*; suoni *semplici* e *doppi*; *zeta dura* o *dolce*) e la scelta di illustrare i suoni seguendo l'ordine alfabetico. Nell'ultima lezione del corso vengono proposti brani in italiano antico e infine viene letto l'inizio dei *Promessi sposi* dapprima con accenti regionali (romano, milanese, calabrese, veneto – invero poco marcati) e da ultimo in italiano standard. La didattica di questo corso si basa essenzialmente sulla grafia: nell'illustrare le vocali, ad esempio, il maestro spiega che "la vocale *a* ha sempre lo stesso suono, aperto", mentre invece "la vocale *e* può avere due suoni, un suono aperto e un suono chiuso e stretto". Vengono enunciate regole senza mai dare alcuna motivazione, al massimo fornendo "norme" meccaniche per la loro applicazione (ad esempio nel caso della conservazione del dittongo mobile). Il secondo corso, a cura del prof. Umberto Pittola, traduttore dall'inglese e docente di fonetica presso l'Università per Stranieri di Perugia, è destinato agli stranieri e sicuramente pensato per un pubblico adulto, dato che procede molto speditamente ed usa una terminologia scientifica per riferirsi ai suoni. Vengono citati come esempi molti brani poetici celebri, accanto a frasi create ad hoc per illustrare un determinato fonema (*È meglio ch'egli non pigli moglie*; *Pippo appena può parte per Padova*). In entrambi i corsi si registra la presenza di termini toscani negli esempi (per esempio *babbo*, *lapis*; viene citato uno stornello come esempio). Al contempo, si notano l'assenza del raddoppiamento fonosintattico, la pronuncia sonora delle sibilanti intervocaliche (laddove l'uso toscano dell'epoca presenta la sorda: *così*; *posare*; *asino*), nonché delle vocali *e*/*o* aperte e chiuse, per le quali viene prescritta talvolta una pronuncia che si discosta dallo standard di base toscano. Queste caratteristiche rivelano la scarsa aderenza tra la corretta pronuncia prescritta dalla norma e la realizzazione della stessa, non solo a livello colloquiale, ma anche nelle proposte istituzionali.

Bibliografia

Camilli, Almerindo & Fiorelli, Piero (1965), *Pronuncia e grafia dell'italiano*, Firenze, Sansoni

Canepari, Luciano (1999[2]), *Il MaPI. Manuale di pronuncia Italiana*, Bologna, Zanichelli (1ª ed. *Manuale di pronuncia italiana, con un pronunciario di oltre 30.000 voci*, 1992)

Corso di ortoepia, a cura del Centro Nazionale Sussidi Audiovisivi; testo di Antonio Mura; dizione di Giovan Battista Arista, Roma, Editrice italiana audiovisivi, 1960

De Mauro, Tullio (1970[2]), *Storia linguistica dell'Italia unita*, Bari, Laterza (1ª ed. 1963)

Galli de' Paratesi, Nora (1984), *Lingua toscana in bocca ambrosiana. Tendenze verso l'italiano standard: un'inchiesta sociolinguistica*, Bologna, il Mulino

Fiorelli, Piero (1964), *Corso di pronunzia italiana*, Padova, Radar

La pronunzia dell'italiano insegnata agli stranieri, a cura di Umberto Pittola, Roma, Editrice italiana audiovisivi, 1961

Malagoli, Giuseppe (1905), *Ortoepia e ortografia italiana moderna*, Milano, Hoepli

Migliorini, Bruno, Tagliavini, Carlo & Fiorelli, Piero (1969), *Dizionario d'ortografia e di pronunzia*, Torino, ERI

Tagliavini, Carlo (1965), *La corretta pronuncia italiana. Corso discografico di fonetica e ortoepia*, Bologna, Capitol, 2 voll.

Language-dependency of /m/ in L1 Dutch and L2 English

Meike M. de Boer & Willemijn F. L. Heeren
Leiden University Centre for Linguistics, Leiden University
{m.m.de.boer, w.f.l.heeren}@hum.leidenuniv.nl

So far, most research in forensic phonetics has been performed in a monolingual context [1]. At the same time, the majority of people are multilingual [2]. Consequently, criminal cases may involve speech samples in multiple languages, sometimes even within one recording [3]. This shows the need to explore the existence of language-independent characteristics within speakers to be used in forensic speaker comparisons. Ideally, such characteristics are highly speaker-specific and are used similarly in the two languages. The current study explores language-dependency of the bilabial nasal /m/ in a group of speakers with L1 Dutch and L2 English. Prior work in a monolingual context has shown that /m/ is among the most speaker-specific segments because of the involvement of the nasal cavity [e.g. 4]. The nasal cavity is relatively rigid when compared to the oral cavity, leading to low within-speaker variability and high between-speaker variability [4]. In addition, in both Dutch and English, /m/ is a common phoneme, which is produced similarly and is used in similar phonetic contexts [5]. Hence, this study investigates whether multilingual speakers may be consistent in their production of /m/ across languages.

Method

In spontaneous monologues of 53 female speakers from D-LUCEA [6], /m/ realizations were investigated. The speakers were L1 speakers of Dutch who learned English as an L2 and had above-average L2 proficiency. They were in their 1st month of undergraduate education at an English liberal arts and science college in The Netherlands. Speakers talked for about two minutes in Dutch and then English about an informal topic of their choice. Tokens were located automatically based on the orthographic transcription and segmented manually in Praat [7]. Tokens were excluded when voiceless, creaky, <30 ms, part of the filled pause *um*, or of a different-language word. Thus, 2,972 /m/ tokens were included (Dutch: 1,681; English: 1,291).

For each token, the following measurements were taken: duration, maximum intensity (iMax), center of gravity (CoG) and its standard deviation (SD), and the first four nasal formants (N1-4) and their bandwidths (BW1-4). To see to what extent speakers' /m/ realizations were language-dependent, linear-mixed effects models were used [8], testing the fixed factor Language (Dutch, English) and random by-speaker slopes for Language. In addition, an indication of within-speaker variability was taken using SDs per speaker.

Results

Results showed that cross-linguistic differences in /m/ acoustics within the same speakers were minor (see table 1). Only for duration and N2, the best-fitting models included Language ($\chi^2(1) = 97.2$, $p < .001$; $\chi^2(1) = 9.56$, $p = .002$). Tokens in L2 English were on average 9 ms longer than those in L1 Dutch. Note, however, that we did not control for speech rate. When looking at spectral characteristics, /m/ tokens in L2 English on average had a 31 Hz higher N2 than in L1 Dutch. The other spectral measurements did not differ across the speakers' languages.

Speakers varied somewhat in the extent to which they made adaptations in the L2: for N2, iMax, CoG, and SD, random by-speaker slopes for Language were included in the best-fitting models. Whereas across speakers, the English N2 was 31 Hz higher, for some individual speakers, it was lower or more similar to the Dutch N2 (see fig. 1).

For all measurements, the means of by-speaker SDs (see table 2) were lower than the SDs across speakers (in table 1), showing that within-speaker variability seems lower than between-speaker variability. Tables 1 and 2 show that SDs were similar in both languages for most measurements, but somewhat larger in English for duration ($t(52) = -3.72$, $p < .001$).

Table 1. Overview of the means (and standard deviations) of the measurements per language.

	L1 Dutch		L2 English			L1 Dutch		L2 English	
Log10_dur (*s*)	−1.22	(0.15)	−1.16	(0.18)	N2 (*Hz*)	1,144	(272)	1,177	(303)
iMax (*dB*)	68	(6)	68	(6)	BW2 (*Hz*)	408	(352)	419	(348)
CoG (*Hz*)	278	(50)	277	(47)	N3 (*Hz*)	2,063	(378)	2,080	(368)
SD (*Hz*)	315	(166)	307	(166)	BW3 (*Hz*)	516	(379)	504	(376)
N1 (*Hz*)	321	(60)	322	(55)	N4 (*Hz*)	2,733	(332)	2,741	(325)
BW1 (*Hz*)	122	(66)	116	(63)	BW4 (*Hz*)	333	(309)	335	(328)

Table 2. Means of the by-speaker SDs in L1

	L1	L2
Log10_dur (*s*)	0.14	0.17
iMax (*dB*)	2.83	2.65
CoG (*Hz*)	40	38
SD (*Hz*)	159	161
N1 (*Hz*)	48	45
BW1 (*Hz*)	59	57
N2 (*Hz*)	253	269
BW2 (*Hz*)	328	318
N3 (*Hz*)	340	323
BW3 (*Hz*)	361	340
N4 (*Hz*)	304	293
BW4 (*Hz*)	283	305

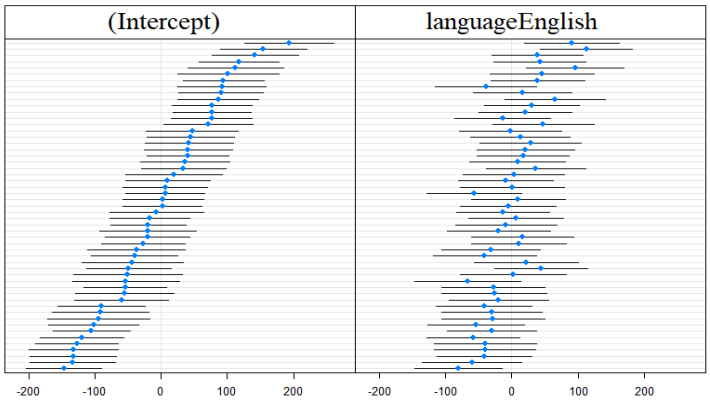


Fig. 1. Caterpillar plot showing the random structure Dutch and L2 English of the N2 model, i.e. by-speaker intercepts (left) and by-speaker adaptations in the L2 (right).

Discussion and conclusion

The acoustics of /m/ seem relatively language-independent within speakers; L1 Dutch speakers showed minimal changes in their /m/ acoustics when speaking in L2 English. The feature showing the clearest cross-linguistic difference was N2, which is associated with the oral and nasal cavities [9]. Hence, despite the rigidness of the nasal cavity, some language-dependent features may remain. Based on these results, /m/ may be a useful segment for cross-linguistic forensic speaker comparisons. However, recording conditions in casework are typically worse than in the data used here, and other L2 speakers may differ in proficiency. Therefore, more research is needed to estimate the strength-of-evidence of /m/ for cross-linguistic casework, and to study /m/ in more or less advanced learners or speakers of different language combinations.

Acknowledgement: This research was supported by an NWO VIDI Grant (276-75-010).

References

[1] Mok, P. P., Xu, R. B., & Zuo, D. (2015). Bilingual speaker identification: Chinese and English. *International Journal of Speech, Language & the Law*, 22(1).

[2] Bhatia, T. K., & Ritchie, W. C. (2012). *The handbook of bilingualism and multilingualism*. West-Sussex, UK: Wiley-Blackwell (pp. xxi – xxiii).

[3] Van der Vloed, D. L., Bouten, J. S., & Van Leeuwen, D. A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. *Proceedings of Odyssey Speaker and Language Recognition Workshop 2014*, Joensuu, Finland, June 16-19, 2014, 6-13.

[4] Rose, P. (2002). Forensic speaker identification. In: J. Robertson (Ed.), *Taylor & Francis Forensic Science Series*. London: Taylor & Francis (pp. 125-173).

[5] Collins, B., & Mees, I. M. (2003). *The phonetics of English and Dutch*. Leiden: Brill (pp. 167-181).

[6] Orr, R., & Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project data. In: J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. Berkeley: Ubiquity Press (pp. 177−190).

[7] Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer [computer program]. Retrieved 3 July 2018 from http://www.praat.org/

[8] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4, *Journal of Statistical Software*, *67*, 1−48.

[9] Fant, G. (1970). *Acoustic theory of speech production* (2nd ed.). The Hague: Mouton.

La variazione prosodica in italiano: l'utilizzo di un chatbot Telegram per la didattica assistita per apprendenti di italiano L2 e nella valutazione linguistica delle conoscenze disciplinari

Valentina De Iacovo*, Marco Palena** e Antonio Romano*
*Università degli Studi di Torino, **Politecnico di Torino

Confrontare la pronuncia di apprendenti di una lingua straniera con enunciati di parlanti nativi (Delmonte 2009) sta ricevendo sempre più attenzione grazie anche alle numerose applicazioni che nascono nell'ambito della didattica assistita. Parallelamente, gli studi glottodidattici sulla variazione prosodica tra più parlanti nativi fanno emergere una variabilità ritmico-intonativa che non può essere ridotta a pochi modelli eligibili ma, al contrario, dovrebbe essere parte del bagaglio linguistico dell'apprendente. Allo stesso tempo, pare ancora difficile in che modo esplicitare all'apprendente il suo grado di competenza prosodica. Seguendo questa direzione, in questo studio presentiamo un *chatbot* pensato come supporto di apprendimento proattivo per il miglioramento delle competenze orali in italiano L2. Realizzato all'interno dell'applicazione di messaggistica istantanea Telegram, il *chatbot* prevede l'interazione con l'utente attraverso domande e risposte basate sulla valutazione di conoscenze disciplinari. In particolare, propone all'apprendente una serie di domande a risposta chiusa (quiz) che possono avere carattere generale di comprensione linguistica oppure essere legate a un particolare ambito disciplinare. All'individuazione della risposta corretta, l'apprendente ha la possibilità di ascoltare la stessa prodotta da parlanti madrelingua. A questo punto, l'apprendente invia al *chatbot* la propria risposta sotto forma di nota vocale. Questo è in grado di confrontare, in maniera automatica, la risposta dell'utente con un archivio di risposte date da parlanti madrelingua e trovare quindi quella che più si avvicina in termini prosodici a quella dell'apprendente. A partire da questa vicinanza prosodica l'utente riceve infine un riscontro sul proprio livello di competenza orale. Inoltre, l'impostazione a quiz permette la valutazione di eventuali criticità linguistiche come ad esempio la pronuncia di date o formule matematiche. Nella seconda parte della presentazione mostriamo quindi i primi risultati di uno studio pilota che vede la partecipazione di parlanti madrelingua italiana (con variazione regionale) e apprendenti di italiano L2. Attraverso il chatbot i locutori hanno risposto a una serie di domande creando così un corpus di frasi che sono state successivamente analizzate. Oltre ai possibili aspetti più tecnici (intensità, velocità d'eloquio, elisioni) ci si è quindi soffermati su quali siano stati i criteri di vicinanza prosodica che hanno portato alla valutazione in percentuale della frase letta dall'utente: ciò ha permesso sia di testare preliminarmente il grado di affidabilità del sistema sia di vedere se la

variazione tra locutori italofoni ampliasse o meno il grado di correlazione con gli apprendenti di italiano.

Bibliografia:

BOUREUX M. & BATINTI A. (2004), La prosodia. Aspetti teorici e metodologici nell'apprendimento-insegnamento delle lingue straniere, in *Atti delle XIV Giornate di Studio del Gruppo di Fonetica Sperimentale*, Esagrafica, Roma: 233-238.

BUSÀ M.G. (2012), The role of prosody in pronunciation teaching: a growing appreciation, in: *Methodological Perspectives on Second Language Prosody* (a cura di M.G. Busà, A. Stella), Padova: Cleup, 101-105.

CAZADE A. (1999), De l'usage des courbes sonores et autres supports graphiques pour aider l'apprenant en langues, in *ALSIC (Apprentissage des Langues et Systèmes d'Information et de Communication, online)*, 2(2): 3-32.

CHUN D. M. (2002), *Discourse Intonation in L2: From theory and research to practice*, Benjamins, Amsterdam.

DELMONTE R. (2009), Prosodic tools for language learning, in *International Journal of Speech Technology* 12(4): 161-184.

FERNOAGĂ V., STELEA GA., GAVRILĂ C. & SANDU F. (2018), Intelligent education assistant powered by chatbots, in *The International Scientific Conference eLearning and Software for Education* 2: 376-383.

LACHERET-DUJOUR A. (2001), Modéliser l'intonation d'une langue. Où commence et où s'arrête l'autonomie du modèle? L'exemple du français parlé, in *Actes du colloque international Journées Prosodie2001*, 57-60.

PEREIRA J. (2016), Leveraging chatbots to improve self-guided learning through conversational quizzes, in *Proceedings of the fourth international conference on technological ecosystems for enhancing multiculturality*, TEEM '16, ACM Press, New York: 911-918.

UNA NUOVA IDEA DI "IMPRONTA VOCALE"

COME STRUMENTO IDENTIFICATIVO E RIABILITATIVO

Autori: Marco Farinella, Marco Carnaroglio, Fabio Cian

Istituto Mod.A.I.® di Torino – segreteria@istitutomodai.it

INTRODUZIONE

Da sempre l'uomo usa il proprio istinto empatico per decifrare i segnali verbali della comunicazione, nel tentativo di riconoscere i propri simili, comprenderne le intenzioni o svelarne le menzogne. L'opinione soggettiva, però, in sede giudiziaria è irrilevante ai fini processuali. Con l'avvento della tecnologia è diventato quindi indispensabile studiare l'*impronta vocale*. Timbro, intonazione, posizione degli accenti, velocità di elocuzione, durata ed epentesi vocalica e consonantica, variazione intra e interlinguistica sono diventati i nuovi campi di ricerca. Numerosi studi hanno indagato tutti questi parametri utilizzando la *stima della frequenza*, i modelli di *misture gaussiane* (GMM) e quelli di *Markov nascosti* (HMM), gli algoritmi di *pattern matching*, le *reti neurali*, le *matrici di rappresentazione*, la *quantizzazione vettoriale*, e gli *alberi di decisione*. Tuttavia, nonostante le varie teorie formulate e le metodiche utilizzate (*cohort model*, *modelli ambientali*, algoritmi di riduzione del rumore, etc…), permangono molti problemi. Le condizioni di ripresa, i sottofondi ambientali, i comportamenti, le inflessioni linguistiche, gli umori, lo stato di salute e l'età del soggetto possono infatti inficiare sia la fase di *raccolta,* sia quella di *verifica* degli elementi, rendendo alquanto controversa la qualità dei risultati ottenuta elettronicamente (Gold e French 2019). In ambito forense ciò rende indispensabile l'affiancamento di esperti con allenamento all'ascolto e conoscenze tecniche specifiche (Romito, Galatà, 2008) affinché l'esito dell'analisi possa ritenersi accettabile in termini probabilistici (Grimaldi, d'Apolito, Gili Fivela, Sigona 2014). Questo contributo propone una diversa e personale interpretazione del fenomeno sonoro durante la verbalizzazione, nell'attesa che la scienza individui nuove metodologie d'integrazione fra le letture strumentali e la percezione umana che possano affiancare, in maniera affidabile, l'indispensabile contributo del tecnico forense.

SCOPO DEL LAVORO

Lo scopo del presente lavoro è approfondire la correlazione fra l'*impronta vocale* e la *fisiologia* che l'ha prodotta, per individuare una *configurazione frequenziale* atta ad identificare e migliorare gli aspetti *funzionali* ed *emotivi* peculiari di una persona.

MATERIALI E METODI

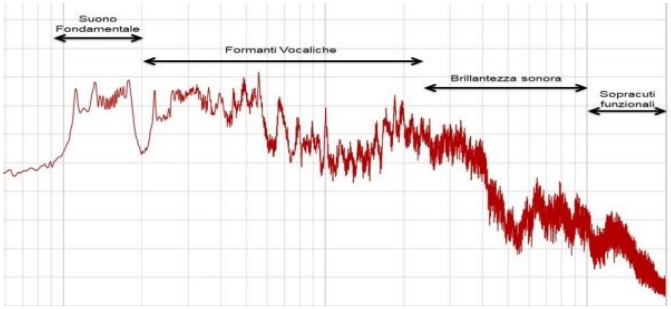


Figura 1: In rosso lo spettro di Fourier di una voce maschile che pronuncia la frase "le mie aiuole". In nero la suddivisione delle aree frequenziali per l'analisi vocale.

Fra il 2003 e il 2020, attraverso un'osservazione funzionale, si è indagato l'atteggiamento laringeo di un campione, composto da 1061 persone (33% uomini e 67% donne), nell'atto della fonazione. Le età medie sono state di 47 anni (fra i 7 e i 91 anni) per i maschi e di 48 anni (fra i 12 e gli 85 anni) per le femmine. Le caratteristiche acustiche delle voci sono state raffrontate con la condizione di salute *fisica* e *psicologica*, con le *abitudini sociali* e i *comportamenti professionali* degli individui esaminati. Per praticità d'analisi l'intera gamma delle *frequenze udibili* dall'uomo è stata suddivisa nei seguenti gruppi (identificativi della regione anatomica in cui si sviluppano): *suono fondamentale* (prodotto dalle pliche vocali), *formanti vocaliche* (proprie della cavità orale), *brillantezza sonora* (influenzata dall'onda della mucosa), *sopracuti funzionali* (frutto della fisiologia degli apparati). Si è proceduto con la tradizionale analisi *prosodico-intonativa* applicandola però a tutti i gruppi, monitorando quindi la derivata di F0 e dell'intera gamma di armonici nelle bande d i maggiore rilevanza. Le indagini sono servite per enucleare un *pattern sonoro* ideale, di perfetto funzionamento laringeo, da attribuire ad un ipotetico adulto, in perfetta salute, scevro da traumi fisici e psicologici, in condizioni emotive di equilibrata serenità. A questo riferimento *teorico* si è attribuito un valore arbitrario, di *affaticamento glottico,* pari allo 0%. Successivamente, utilizzando un passaggio d'aria durante il meccanismo di rilassamento della *chiusura glottale* (Greene e Mathieson, 2001) è stata misurata la risonanza laringea normale delle laringi a riposo. Comparando statisticamente le *trasformate di Fourier* (FFT) reali mediate dell'intera frase con il *modello ideale* si sono evidenziate le discordanze dovute alle costrizioni laringee tipiche del soggetto considerato ed attribuito a ciascuno un valore di *affaticamento* vocale espresso in percentuale. Tale punteggio rappresentava la condizione fonatoria quotidiana dell'individuo, nel periodo osservato, generando anche una nuova concezione di "impronta vocale" più vicina a fattori biologici che ad elementi linguistici (Lindh, 2004). Infine, per valutare la qualità dei dati raccolti e la bontà delle conclusioni, si è deciso di testare anche il modello di riferimento, invertendo l'ottica d'indagine. Fra il 2005 e il 2011, 171 volontari sono stati sottoposti ad un'unica stimolazione percettiva che inducesse la loro voce ad assomigliare al *pattern sonoro ideale*. L'idea era di elicitare sensorialmente le voci verso un *punto zero fisiologico di affaticamento* e sondarne la retroazione, funzionale ed emotiva, impattante sulla percezione del sé, sull'identità e sul conseguente modello comportamentale. Per la valutazione è stato usato un test con venticinque coppie di aggettivi contrapposti (ispirato a Costa e McCrae, 1992), disposti in maniera casuale, che gli esaminandi compilavano prima e dopo la pratica di vocalizzazione.

RISULTATI

Paragonando l'intero spettro sonoro delle voci con il FFT ideale si è determinato il grado di funzionalità delle prestazioni vocali senza ricorrere ai tradizionali esami clinici che, per la loro complessità, sono eseguibili soltanto in apposite strutture sanitarie con personale qualificato. Ciò, di fatto, ha snellito notevolmente la procedura. Inoltre si è evidenziato come la percentuale di discostamento dal *modello di riferimento* tenda a mantenersi costante per ogni esaminato a parità di condizioni psicofisiche, caratterizzando quindi l'oratore, pur non essendo ancora riusciti a parametrizzare la lettura strumentale per il riconoscimento univoco del parlante senza l'intervento umano. Tuttavia, il risultato certamente più sorprendente è stato ottenuto utilizzando il *pattern vocale ideale* come stimolazione per il miglioramento della prestazione fonatoria. Infatti, l'esito dell'esperimento ha messo in luce la sua grande capacità rieducativa (priva di effetti collaterali) e la forte retroazione sull'identità dei volontari stimolati. L'affaticamento glottico si è notevolmente ridotto ed i test psicologici hanno conseguentemente evidenziato una variazione drastica del quadro emotivo, con due importanti traguardi: l'eliminazione dei vocaboli con un'accezione negativa nella descrizione del proprio stato psicofisico (ad es. teso, pesante, stanco, triste…) ed un incremento della propriocezione dei soggetti verso gli aggettivi positivi (attivo, riposato, allegro, concentrato…) con percentuali di miglioramento che vanno dal 6% fino al 53%, con una media globale di poco inferiore al 30% ottenuta in un'unica seduta di stimolazione.

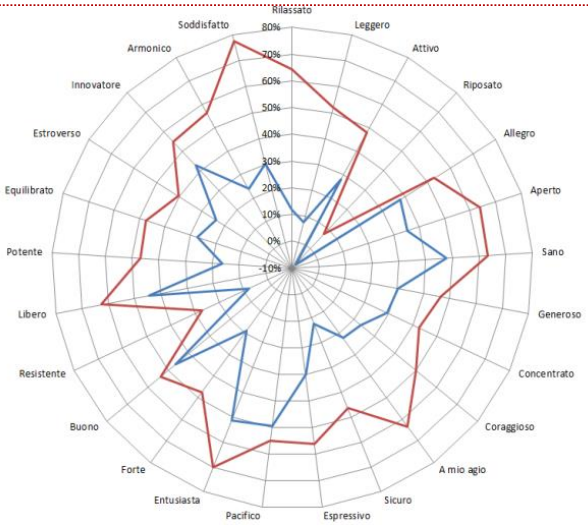


Figura 2: In blu la condizione emotiva dei soggetti prima della stimolazione, in rosso il cambiamento propriocettivo al termine della seduta

CONCLUSIONI

La voce, nel ventaglio delle sue ventimila frequenze e grazie alle oltre ventisei milioni d'informazioni empatiche al secondo che esse portano con sé, è certamente ascrivibile all'identità individuale della persona. La muscolatura del tratto vocale reagisce istintivamente alle emozioni come primaria difesa dei polmoni. Questo fa sì che la risonanza di ogni persona sia diversa e assimilabile ad una impronta digitale. Il vissuto di un individuo resta registrato nella sua biologia, caratterizzandone la personalità emotiva, comportamentale e quindi anche quella sonora globale. Per esaminare la condizione *psicofisica* di un individuo si ritiene pertanto necessario considerare l'intera configurazione frequenziale, valutando il differenziale della sua *impronta vocale mediata di una intera asserzione* (intesa nell'accezione del presente studio) da una *trasformata di Fourier ideale*. Perfezionando questa tecnica, non si esclude una futura applicazione in ambito forense, non tanto per validare le intercettazioni (normalmente troppo disturbate da rumori parassiti che interferiscono con l'esame audio), quanto per valutare il grado di *stress laringeo* durante una dichiarazione giurata. L'incremento di tensione muscolare nel pronunciare una frase (rilevato paragonando l'*analisi della voce* durante una deposizione con la propria *impronta laringea*) potrebbe risultare assai prezioso, come parametro oggettivo, per valutare la veridicità di una dichiarazione. Tuttavia, in attesa di creare una procedura automatizzata e sicura per l'osservazione tecnica ipotizzata, questa nuova concezione di *impronta vocale ideale* può già essere utilizzata con successo come strumento riabilitativo, sia dal punto di vista fisiologico, sia psicologico. Infatti, quando è impiegata come *modello acustico,* all'interno di una *pedagogia sensorialmente orientativa*, essa è in grado di *rieducare* velocemente una funzione glottica deficitaria e di retroagire fortemente sull'umore, con effetti duraturi di miglioramento sulla percezione del sé. La voce così riabilitata aumenta il suo grado di funzionalità fisiologica e l'identità delle persone trattate risulta più definita, equilibrata e risoluta.

BIBLIOGRAFIA

- Costa P.T., McCrae R.R., in *Psychological Assessment Resources* (1992)
- Federico A., Paoloni A., in *Media Duemila 250* (2007)
- Gold E., French P., in *International Journal of Speech Language and the Law* (2019)
- Greene M. e Mathieson L., "The Voice and its Disorders" (2001)
- Grimaldi M. et al., in *Mondo digitale* (2014)
- Hamidi M. et al., in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (2020)
- Kelly F. et al., in *Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics* (2019)
- Lindh J., "Handling the Voiceprint Issue" (2004)
- Romito L., Galatà V., in *Language Design*, vol. Special issue I, (2008)
- Saleem S. et al., in *Forensic Science International: Digital Investigation* (2020)
- Zetterholm E., "Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success", PhD thesis, Lund University. (2003)

Chloë Farr, Gracellia Purnomo, Amanda Cardoso, Arian Shamei and Bryan Gick

Speaker Accommodations and VUI Voices: Does Human-likeness of a Voice Matter?

Voice user interfaces (VUI) are becoming increasingly embedded in peoples' lives, as they are built into every voice assistant (e.g. Amazon's Alexa, Apple's Siri and Microsoft's Cortana) on smartphones, computers, and smart home products (Amazon Echo, Google Home), and are able to perform a large number of automated tasks. Therefore, people may be speaking to VUI with increasing ease and frequency [Rubio-Drosdov, E., Díaz-Sánchez, D., Almenárez, F., Arias-Cabarcos, P., & Marín, A. (2017). Seamless human-device interaction in the internet of things. *IEEE Transactions on Consumer Electronics*, *63*(4), 490–498]. Such increases may have an effect on the way that people perceive and interact with the devices, so that interactions with VUI become more like those with human interlocutors (human-to-animate), such as adjusting the properties of one's speech to be better understood (e.g. hyperarticulation) or to be more or less like the interlocutor (i.e. accommodation). For example, previous work has demonstrated that speakers may attempt to produce clearer or more understandable speech by changing phonetic properties of their speech, such as changing their pitch (f0) or making their voice louder [Oviatt, S., Maceachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*, 87-110]. These types of phonetic adjustments have also been found as a communication repair strategy used in human-to-animate interactions and have often been described under the label of hyperarticulation [Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. *Speech Production and Speech Modeling*, *55*, 403–439]. This might suggest that human interaction with VUI are, at least, in some ways like human-to-animate interaction. However, much of the current work looking at the phonetic changes that are observed in human-to-VUI interactions was conducted at a time when VUI were newer, not as advanced in terms of uses or the way they sound, and less integrated into devices and humans' lives. The limited more recent work has generally been impressionistic and focused on how understanding phonetic changes can enhance automatic speech recognition systems (ASR) rather than what the implications are for speakers and changes in their speech behavior. Understanding more about the phonetic characteristics of human speech in human-to-VUI interaction may shed light on people's expectations of the cognitive capabilities and animacy of the devices and also the changing nature of human-to-VUI interactions, as they become more intertwined in our daily lives. Our wider project is interested in investigating whether people treat VUI as animate (more human-like) or inanimate technology (more like a tool). The current paper focuses on one part of this question by looking at whether the human-likeness of VUI's synthetic voice (text to speech - TTS) influences the extent that speaker accommodation occurs for the human interlocutor, and what this can tell us about how speakers categorise VUI and, in turn, VUI interaction as either different or the same as other spoken interactions.

A large body of research demonstrates that speakers adjust the properties of their speech to be more or less like their interlocutor depending on whether the speaker perceives the interlocutor to be part of their same in-group (e.g. from the same region, community, ethnic background, socio-economic status, etc.) or not (for work on accommodation, see Giles 1973, Pardo 2006, Giles & Ogay 2007, Babel 2012). Speakers converge (sound more like) to interlocutors that they consider to be in their group or who they would like to lessen social distance with, and they diverge (sound less like) from interlocutors that they consider to be outside of their group or with whom they would like to increase social distance from themselves. We, therefore, hypothesise that if the speaker considers a VUI voice to be in-group (i.e. group them as like a human interlocutor) then they will converge to the VUI's speech characteristics and if they do not they will either diverge from the VUI or demonstrate maintenance (i.e. no accommodation). Whether they diverge or show maintenance will also provide us with an understanding of the way that human participants perceive these voices. Divergence would specifically suggest an "othering" and wanting to create distance, demonstrating superiority/power or categorising the VUI voice as specifically non-human. On the other hand, maintenance does not necessarily suggest that the participant is distancing themselves, but rather that this is not comparable to human-to-animate interactions. As the exposure to VUI and interactions have increased most substantially over the last decade, it might also be the case that

speakers would interact with any VUI or no VUI in a similar way to human-to-animate interactions. In the current study we investigate these hypotheses by varying how human-like the VUI's synthetic voice sounds.

We will present an expected 50 participants with two female sounding synthetic voices generated with Amazon Polly's text-to-speech service (Amazon Web Services, 2020), which participants are told are VUI. 26 linguistic researchers rated the two voices as the most human-like and most robotic, while still being perceivable as synthetic voices. One voice uses Polly's Neural system (a sequence of phonemes converted to a sequence of spectrograms, subsequently converted into continuous audio signal) or Polly's Standard system (concatenated pre-recorded phonemes). In order to be able to explore how speakers accommodate to the different voices, we manipulated word-initial voiceless plosive voice onset time (VOT) duration in Praat (Boersma & Weenink, 2020), so that it's twice as long for the robotic voice and half as long as for the more human-like voice compared with the original samples. We also normalised the voices for speech rate, as it was found that there were substantial differences in speaking rate across the robotic and human-like voices.

The experiment is counterbalanced for presentation of voices to participants, so that half of the participants are presented with the more robotic voice first and the other half are presented with the more human-like voice first, and participants are randomly assigned to hear one of these voice orders. Participants are asked to use ten pre-scripted prompts with the VUI, the order of which is randomised within a block. The VUI responds to each of the prompts appropriately. For example, the participant speaks the prompt, "Send a text to Paul", to which the synthetic voice returns, "What would you like to write to Paul?".

Participants' speech are analysed acoustically within each block in order to determine whether VOT duration of word-initial voiceless plosives for the participants are adjusted over the course of the interaction with each of the VUI voices. In other words, is there evidence of accommodation by the participants to the speech characteristics of the VUI's voices and the magnitude of the change that occurs? Preliminary results suggest that half of the participants accommodate to the human-like voice in comparison to the pre-exposed and robotic data, resulting in statistically significant shorter VOT ($p<0.05$). In all participants, VOT accommodation did not occur in relation to the robotic voice. A survey administered after the study reported that the participants believed they were speaking to real virtual assistants.

As previously mentioned, this preliminary study feeds into a wider project that investigates the phonetic properties of speech during human-to-VUI interaction, adjustments to speech that occur as a result of differences in the backgrounds of the human interlocutor (e.g. familiarity with VUI, language variety) and differences in the properties of the VUI (e.g. accuracy of responses, language variety), and how these changes relate to those seen in human-to-human interactions. Ultimately, this project will provide us with insights into under what conditions, if any, humans interact with these devices in the same way they do with other humans, which may give us a better understanding of how these devices are perceived. Finally, understanding if TTS systems mold human speech. If people are accommodating to the VUI, and the VUI is learning from their speech, will ASR natural speech recognition improve, or will there be a fundamental change in human speech production with technological or animate interlocutors?

Radici identitarie e mantenimento linguistico. Il caso di un gruppo di *heritage speakers* di origine calabrese

Manuela Frontera, Università degli Studi Internazionali di Roma

Obiettivi

Il presente lavoro si propone l'obiettivo di misurare il peso esercitato da due variabili psico-sociali (a. senso di appartenenza alla cultura ereditaria; b. integrazione nel paese ospitante) sul mantenimento di specifici tratti linguistici nella L1 di un gruppo di *heritage speakers* calabresi residenti in Argentina. Scopo ultimo è quello di valutare se e in che modo un ipotetico attaccamento e senso di appartenenza alla comunità d'origine possa correlarsi o meno alla produzione linguistica del gruppo di indagine, in relazione ai livelli di mantenimento di un tratto marcato e tendenzialmente stigmatizzato, come la post-aspirazione dei suoni occlusivi sordi nelle varietà calabresi (Nodari, 2015). Inoltre, saranno altrettanto esaminati abitudini e usi linguistici del gruppo indagato, al fine di verificare una possibile interazione fra tali fattori e variabili di tipo psico-sociale.

Introduzione e stato dell'arte

La ricerca si inserisce all'interno di un progetto di più ampio respiro, rivolto all'indagine sociolinguistica di un gruppo di emigrati di prima generazione, provenienti dalla terza aerea dialettologica della Calabria (vd. Trumper, 1997) e stanziati, a partire dal secondo dopoguerra, nelle province argentine di Córdoba, Santa Fe e Buenos Aires, dunque *heritage speakers* di varietà dialettali, a contatto prolungato con la varietà spagnola del paese d'accoglienza. I primi risultati della ricerca intrapresa hanno portato a confermare l'ipotesi di attrito fonetico in atto, inteso come modifica parziale o totale dei tratti della L1, in condizioni non patologiche, a seguito dell'interazione con una nuova lingua in età post-adolescenziale e in contesto migratorio prolungato[1]. In particolare, ad essere indagato è stato il parametro di aspirazione (in termini di Voice Onset Time) nella produzione delle consonanti occlusive sorde /p t k/ in posizione post-sonorante (post-nasale e post-liquida): tendenzialmente *long lag* nelle varietà di origine (vd. Frontera, 2018; Frontera, Tarasi, Graziano, 2019), ma *short* nella varietà spagnola di contatto (Borzone & Guerlekian, 1980; Soto-Barba & Valdivieso, 1999), il ritardo nell'attacco della sonorità manifestato dai parlanti è sembrato posizionarsi su durate intermedie fra i sistemi di riferimento (vd. Frontera, 2020), in linea con l'idea della comparsa di un *midpoint system* (Flege, 1987, 1995) nato dalla pressione esercitata da una nuova lingua acquisita in età adulta su una varietà nativa ormai vulnerabile (vd. ad esempio de Leeuw, 2019; Major, 1992).

Studi condotti su alcune comunità di emigrati a Toronto hanno dimostrato, inoltre, come il comportamento linguistico di parlanti di una lingua ereditaria possa, in qualche misura, essere condizionato dall'attitudine sviluppata dai suddetti parlanti nei confronti della varietà linguistica e, per esteso, della cultura e del paese d'origine (Nagy, 2015; Nagy & Kochetov, 2013; Nodari, Celata, Nagy, 2019). Nello specifico, Nagy ha individuato nel proprio EOQ (*Ethnic Orientation Questionnaire*) uno strumento utile a quantificare degli atteggiamenti impliciti, da poter correlare numericamente alle misure acustiche estrapolate da analisi sul parlato (vd. Nagy, Chociej, Hoffman, 2014). Per rispondere all'obiettivo dato, in questo lavoro si proporrà un riadattamento dell'EOQ al caso della comunità calabrese qui indagata, i cui esiti saranno valutati, quantitativamente e qualitativamente, in rapporto alle durate di VOT prodotte dallo stesso gruppo di parlanti e misurate nelle indagini precedenti (vd. Frontera, 2020).

Metodologia di ricerca

Il campione d'analisi è costituito da 10 emigrati trilingue: la loro L1 è una varietà dialettale del paese d'origine, afferente alla terza area dialettologica della Calabria; la L2 è l'italiano appreso a scuola prima dell'esperienza migratoria; la L3 è lo spagnolo d'Argentina, acquisito spontaneamente dopo l'emigrazione. I soggetti sono equamente bilanciati per sesso (5 donne + 5 uomini), hanno un'età media di 80 anni e risiedono stabilmente in Argentina, mediamente da 65 anni.

I dati sono estratti sottoponendo ciascun soggetto partecipante a un'intervista semi-guidata, costituita da 50 domande (*items*) volte a estrapolare tre diversi indicatori:

1. abitudini e usi linguistici relativi alle tre varietà di riferimento (12 domande);
2. atteggiamento verso la varietà linguistica ereditaria (20 domande);
3. integrazione nella cultura ospitante (18 domande).

In funzione di una prima analisi quantitativa, le risposte ottenute in forma ordinale (es. *sì – a volte – mai*; *molto – poco – per niente*, vd. Nagy et al., 2014) sono convertite in variabili cardinali e, dunque, trasformate in punteggi (da 0 a 2). Nel rispetto della condizione numerica, secondo cui è possibile sommare fra loro soltanto variabili con uguale o simile estensione di scala (Marradi, 2007), è stata adottata la stessa scala per tutte le variabili e per i tre indicatori.

[1] cfr. de Leeuw (2019); Major (1992); Schmid, Köpke (2013).

Ancora, all'interno di ciascun indicatore tutte le scale sono state orientate nella stessa direzione, invertendo i punteggi delle variabili con orientamento semantico differente. Per gli indicatori 1 e 2, il valore massimo della scala è associato, rispettivamente, a un uso maggiore/più diffuso e a un atteggiamento positivo nei confronti della varietà dialettale di provenienza e, nel secondo caso, della cultura di origine. Di contro, uno scarso livello di integrazione nella comunità linguistico-culturale del paese ospitante viene associato al minimo valore della scala di valutazione, per cui l'indicatore 3 ha orientamento opposto rispetto ai precedenti. I dati sono poi normalizzati al fine di ottenere punteggi comparabili in riferimento a ciascun indicatore e ogni informante coinvolto/a. Verificata l'attendibilità e la coerenza interna alle risposte tramite test α di Cronbach (valori ≥ 0.60), gli indicatori sono utilizzati come elementi di correlazione e fattori fissi tramite cui ispezionare statisticamente le variazioni nelle durate di VOT prodotte nella lingua ereditaria[2]. I risultati sono esaminati in relazione alla totalità del gruppo e a ogni singolo/a partecipante. Le risposte vengono poi analizzate e interpretate qualitativamente, allo scopo di verificare e rafforzare quanto emerso dai dati statistici.

Risultati preliminari

Un'osservazione preliminare dei risultati relativi alle analisi sui questionari e alla correlazione di questi ultimi coi dati acustici, suggerisce: i) la quasi totalità dei parlanti italo-argentini esibisce indici medio-alti inerenti al senso di attaccamento verso le proprie origini culturali e linguistiche (a partire da 1,4/2); ii) ciononostante, l'uso della L3 sembra aver soppiantato interamente l'utilizzo della varietà dialettale di origine (con indici sull'uso della L1 prossimi allo 0/2), e gli stessi intervistati manifestano uno spiccato livello di integrazione nella comunità linguistico-culturale del paese ospitante (indici prossimi all'1,8/2); iii) di conseguenza, l'indicatore di abitudini linguistiche appare avere una correlazione positiva con l'abbassamento dei valori di VOT riscontrati nei parlanti; tuttavia, si osserva correlazione significativa fra alti indicatori attitudinali e più alti livelli di aspirazione. Questi dati preliminari saranno confermati e arricchiti dalle analisi ancora in corso.

Riferimenti bibliografici

Borzone, A. M., Guerlekian, J. (1980). Rasgos acústicos de las consonantes oclusivas españolas. In *Fonoaudiología*, 26 (3): 326-330.

de Leeuw, E. (2019). Phonetic attrition. In Schmid, M. S. and Köpke, B. (Eds.), *The Oxford Handbook of Language Attrition*, Oxford: Oxford Unversity Press, 202-217.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. In *Journal of Phonetics*, 15, 47-65.

Flege, J. E. (1995). Second Language Speech Learning: Theories, Findings and Problems. In Strange, W. (Ed.), *Speech perception and linguistic experience: issues in cross-language research*, Timonium, MD: York press, 233-277.

Frontera, M. (2018). Aspirated voiceless stops in elderly speakers from Calabria: a pilot study. In Botinis, A. (Ed.), *Proceedings of the 9th Tutorial and Research Workshop on Experimental Linguistics*, Paris, France, 28-30 August 2018, 33-36.

Frontera, M. (2020). Assessing first language phonetic attrition in Italian-Argentinian migrants, poster presentato al XVI Convegno Nazionale AISV, *La variazione linguistica in condizioni di contatto: contesti acquisizionali, lingue, dialetti e minoranze in Italia e nel mondo*, Università della Calabria, 29-31 gennaio 2020, Rende (CS), Italia.

Frontera, M., Tarasi, A. and Graziano, E. (2019). Le consonanti occlusive sorde aspirate in Calabria: un confronto tra aree dialettali. In Calamai, S., Piccardi, D. E Ardolino, F. (Eds.), *Gli archivi sonori al crocevia tra scienze fonetiche. Informatica umanistica e patrimonio digitale*, Milano: Officinaventuno, 293-307.

Major, R. C. (1992). Losing English as a First Language. In *The Modern Language Journal*, 76 (2): 190-208.

Marradi, A. (2007). *Metodologia delle scienze sociali*, Bologna: Il Mulino.

Nagy, N. (2015). A sociolinguistic view of null subjects and VOT in Toronto heritage languages. In *Lingua*, 164 (2), 309-327.

Nagy, N., Chociej, J. & Hoffman, M. F. (2014). Analyzing Ethnic Orientation in the quantitative sociolinguistic paradigm. In *Language & Communication*, 35: 9-26.

Nagy, N. & Kochetov, A. (2013). Voice onset time across the generations: A cross-linguistic study of contact-induced change. In Siemund, P., Gogolin, I., Schulz, M.E., Davydova, J. (Eds.), *Multilingualism and language contact in urban areas: Acquisition—Development—Teaching—Communication*, Amsterdam: John Benjamins Publishing, 19-38.

Nodari, R. (2015). Descrizione acustica delle occlusive sorde aspirate: analisi sociofonetica dell'italiano regionale di adolescenti calabresi, in Vayra, M., Avesani, C. & Tamburini, F. (Eds.), Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio, Studi AISV 1, Milano: Officinaventuno, 139-153.

Nodari, R., Celata, C. & Nagy, N. (2019). Socio-indexical phonetic features in the heritage language context: Voiceless stop aspiration in the Calabrian community in Toronto. In *Journal of Phonetics*, 73, 91-112.

Schmid, M.S., Köpke, B. (2013). *First Language Attrition*, Amsterdam: John Benjamins Publishing.

Soto-Barba, J. and Valdivieso, H. (1999). Caracterización fonético-acústica de la serie de consonantes /p-t-k/ vs. /b-d-g/. In *Onomazein*, 4: 125-133.

Trumper, J. (1997). Calabria and southern Basilicata. In Maiden M. and Parry M. (Eds.), *The dialects of Italy*, London: Routledge, 355-364.

[2] La metodologia d'analisi adottata e i risultati ottenuti per i dati acustici di riferimento sono dettagliati in Frontera (2020).

Modeling intonation in interaction.
A new approach to the intonational analysis of questions in (semi-)spontaneous speech

Davide Garassino[1], Dalila Dipino[1] & Francesco Cangemi[2]
([1] Institute of Romance Studies, University of Zurich; [2] Institute of Linguistics – Phonetics, University of Cologne)

The aim of this contribution is to present a new approach to the analysis of intonation in (semi-)spontaneous conversational settings, as a complement to the currently widespread analysis of read speech. We explore the idea that studying intonation in a rich pragmatic and interactional context yields precious insights, and we conclude that the benefits of meaningful communicative contexts offset the drawbacks imposed by challenging phonetic analyses.

The data that we discuss were gathered from a *Map Task* administered to 16 native speakers of Genoese (average age= 63; 6 females), a Northern Italo-Romance variety (cf. Forner 1988). Even if a full-fledged intonational description of this dialect is not currently available, a preliminary analysis of Genoese declarative and interrogative sentences is offered by De Iacovo (2017: 54-55) (for recent intonational analyses of the regional Italian variety spoken in Genoa, see Crocco 2011 and Savino 2012). Our *Map Task* was inspired by the *Montclair Map Task*, proposed by Pardo et al. (2019). Following this model, we have cancelled the 'classic' instruction Giver's and Follower's roles, in order to provide a more natural interactional setting. The speakers were provided with two maps, in which they were supposed to follow an already drawn path and which presented both shared and unshared landmarks (the target words were minimal pairs originally used for gathering data on the phonetic realization of short and long vowels. Since this is not the topic at issue here, we will not discuss it any further). The speakers were instructed to compare the items on their maps in order to 'spot the differences' (this is a feature of the *Montclair Map Task* that makes it similar to the 'spot the difference' task already used in the CLIPS corpus, cf. Albano Leoni & Giordano 2005, and in the *Diapix*, cf. van Engen et al. 2011; Baker & Hazan 2011). Recordings were automatically annotated into InterPausal Units using temporal criteria (i.e. pauses of at least 200ms). IPUs were then manually verified for boundary position. Out of the 1527 interpausal units composing the corpus, 72 were coded as questions. We excluded a large number of items for phonetic (e.g. inaudible overlaps), structural (e.g. false starts) and functional (e.g. stylized mockery) reasons. The 41 remaining items were judged reliable enough to undergo further analysis. Questions were then annotated with respect to their logical semantic structure (polar, alternative, *wh*-questions) and discourse-pragmatic functions (information-seeking, check, rhetorical), cf. Enfield et al. 2010 and Stivers & Enfield 2010, as well as their possible epistemic biases (biased, unbiased).

Our qualitative analysis has allowed us to recognize some recurring patterns in Genoese questions, such as a steep fall on the last stressed syllable, followed by a final rise, and often preceded by a high pitch region. Savino (2012: 34-35) also documents fall-rises in Genoese Italian, especially for post-nuclear material in polar questions. Note that this pattern was not always recognized as such by the Praat in-built pitch detector, due to the intrinsically "noisier" nature of conversational speech (Figure 1). This problem was solved by manually verifying pitch points as requested by the creation of periograms (Figure 2), where traditional f0 contours are modulated by periodic energy (Albert et al. 2018; Cangemi et al. 2019). In these displays, the curve "is wide and solid at the most periodic portions and it becomes gradually narrower and more transparent as periodic energy drops" (Cangemi et al. 2019: 807), with red portions indicating interpolated material. On the one hand, the absence of the final rise in the off-the-shelf Praat pitch extractor confirms the necessity for high-quality and reliable phonetic representations in the study of conversation. On the other hand, the possibility of extracting reliable and readable f0 representation from spontaneous data confirms that the ease of treatment of carefully elicited monological speech should not be prized beyond reason.

Crucially, in our corpus the steep fall with low rise was present in ca. 39% of the data for polar (Figures 2 and 4), alternative (Figure 3) and *wh*-questions, as well as for information-seeking (Figures 2 and 3) and check (Figure 4) functions. Therefore, it might seem that this pattern is used in Genoese independently of speakers and semantic types of questions as well as, from a discourse-pragmatic perspective, in both information-seeking and check contexts. This pattern does not seem to be attested in rhetorical questions (Figure 5), or in exchanges in which one speaker seems to perform not only a request for information, but also to express other types of meaning (such as specific attitudes and emotions related to the conversational dynamics). This is the case of the item in Figure 6, in which the speaker is asking for the fifth time about the possible presence of a [fryːtu] 'fruit' landmark on the interlocutor's map, after having failed to receive a satisfactory answer to his previous questions.

These cases, anecdotal as they might seem, suggest that the analysis of conversational speech holds great promise for the understanding of intonation. They raise relevant issues for the study of prosody, such as (i) whether question intonation is as frequent in conversation as its high rate of representation in read speech research might lead us to believe; (ii) whether new visualization and representation techniques can make possible the widespread use of conversational speech in intonation research; and (iii) whether question intonation in interaction might be more sensitive to conversational functions than to well-studied semantic types and pragmatic properties, thus suggesting different directions for the future of intonation studies.

Speaker RM

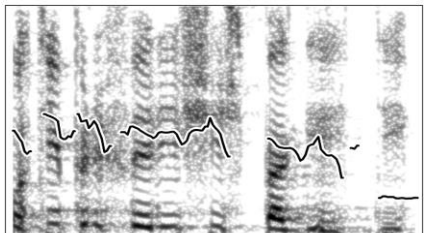


Figure 1. Praat - Polar Question, Information-seeking. *O ti preferisci i spaghetti?* "Or do you prefer spaghetti?"

Speaker RM

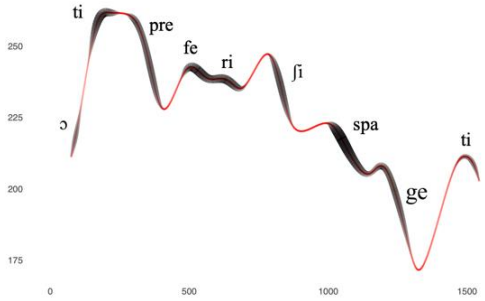


Figure 2. Periogram - Polar Question, Information-seeking. *O ti preferisci i spaghetti?* "Or do you prefer spaghetti?"

Speaker CG

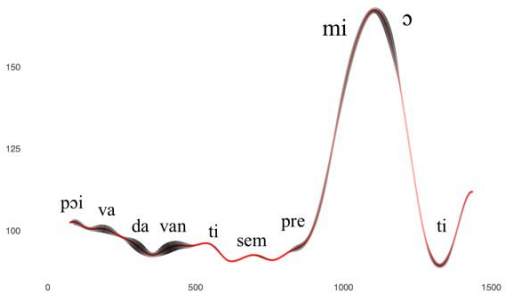


Figure 3. Alternative Question, Information-seeking. *[Poi vado avanti] sempre mi o ti?* "Then do I go on, always me or you?"

Speaker BG

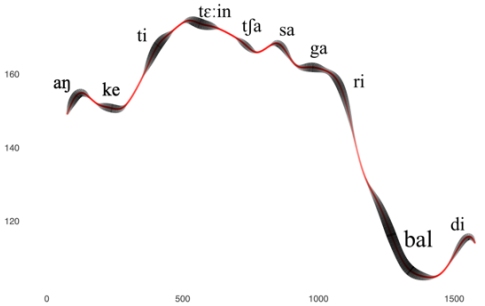


Figure 4. Polar Question, Check. *Anche ti t'è in Ciassa Garibaldi?* "Are you in Garibaldi Square too?"

Speaker GM

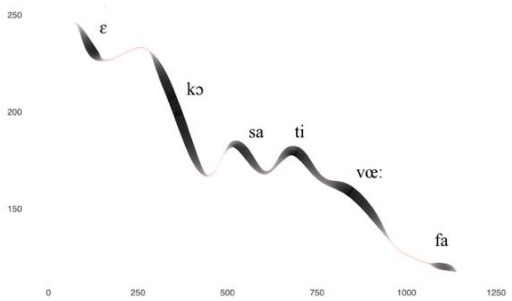


Figure 5. *Wh*-question, Rhetorical. *Eh, cosa ti voe fa?* "Eh, what do you want to do?"

Speaker BG

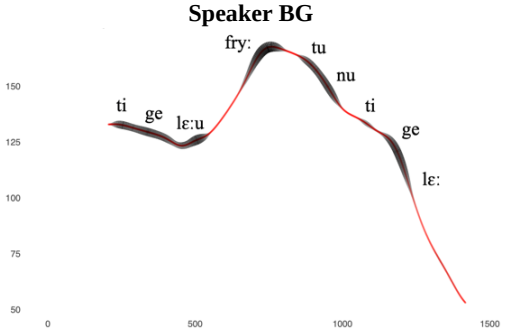


Figure 6. Alternative Question, Check. *Ti ghe l'è u frytu o nu ti ghe l'è?* "Do you have the fruit, or don't you have it?"

References

Albano Leoni, F. & Giordano, R. (eds.) (2005). *Italiano parlato. Analisi di un dialogo*. Napoli: Liguori Editore.

Albert, A., Cangemi, F. & Grice, M. (2018). Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proc. 9th International Conference on Speech Prosody*, 804-808.

Baker, R. & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, *43*(3), 761-770.

Cangemi, F., Albert, A. & Grice, M. (2019). Modelling intonation: Beyond segments and tonal targets. In *Proc. 19th International Congress of Phonetic Sciences*, 572-576.

Crocco, C. (2011). Profili melodici della varietà genovese. In *Atti del VII convegno dell'Associazione Italiana di Scienze della Voce*. Bulzoni Roma,188-199.

De Iacovo, V. (2017). *Intonation analysis on some samples of Italian dialects: an instrumental approach*. Doctoral thesis, Università degli Studi di Genova e Torino.

Enfield, N. et al. (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42 (2010), 2615-2619.

Forner, W. (1988). Areallinguistik I: Ligurien. In Holtus, G., Metzeltin, M. & Schmitt, C. (eds.), *Romanistischen Linguistik (LRL)*, vol. IV, Tübingen: Niemeyer, 453-469.

Pardo, J. S et al. (2019). The Montclair map task: Balance, efficacy, and efficiency in conversational interaction. *Language and speech*, *62*(2), 378-398.

Savino, M. (2012). The intonation of polar questions in Italian: Where is the rise?. *Journal of the International Phonetic Association*, *42*(1), 23-48.

Stivers, T., Enfield, N. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics* 42(10), 2620-2626.

Van Engen, K. J. et al. (2010). The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and speech*, 53(4), 510-540.

GLENDA GURRADO[1]

[1]Università degli studi di Bari

Sulla codifica e decodifica della sorpresa

Il concetto di senso di sorpresa è solitamente associato all'idea di cogliere qualcuno impreparato attraverso un'azione inaspettata, che desta stupore. La sorpresa è annoverata fra le emozioni base tra cui rientrano anche gioia, tristezza, paura, disgusto e rabbia[1]. Le emozioni fondamentali sono selezionate in base ai dati relativi al riconoscimento universale delle emozioni per mezzo delle espressioni facciali e vocali (tra gli altri, Tomkins, 1962, 1963; Izard, 1994; Ekman, 2003). Alcune emozioni prediligono un mezzo anziché un altro a causa del diverso modo di reagire agli stimoli esterni del nostro organismo; tali reazioni si consolidano a livello culturale e sociale, portando a una cristallizzazione delle associazioni fra emozioni e risposte fisiologiche. Si sostiene, dunque, che le emozioni che stimolano un certo tipo di reazioni a un certo grado sono accumunate dall'attivazione dei medesimi pattern prosodici (Scherer, 1986). A seconda del livello di attivazione delle alterazioni fisiologiche che stimolano, si distingue fra emozioni ad alta attivazione ed emozioni a bassa attivazione: le prime (paura, gioia, sorpresa e rabbia) sono generalmente caratterizzate da alti valori di F0, un'estensione tonale ampia, intensità alta, una maggiore velocità di eloquio e quindi una riduzione del numero di pause; diversamente, le seconde sono veicolate da bassi valori di F0, un *range* tonale compresso, intensità bassa, un eloquio più lento e lunghe pause[2]. Nella maggior parte degli studi dedicati ai correlati acustici delle emozioni, la sorpresa è meno indagata rispetto alle altre e, in alcuni casi, addirittura esclusa (fra gli altri, Juslin & Laukka, 2003; Scherer, 2003). Secondo alcuni autori, questa emozione, assieme al disgusto, è difficilmente identificata tramite la voce (Scherer, 1989): probabilmente, nel corso dell'evoluzione la difficoltà di comunicare queste emozioni ha stimolato un tipo di espressione per lo più connessa alla mimica e meno al coinvolgimento delle risorse vocali (Jonhstone & Scherer, 2000). Tuttavia, contrariamente a questa teoria, in alcuni contributi sperimentali l'emozione della sorpresa è riconosciuta dagli uditori in buona percentuale (per l'inglese Cahn, 1990; per lo svedese, Abelin & Allwood, 2000; per lo spagnolo, Iriondo *et al*., 2000). In alcuni contesti la sorpresa è stata, però, confusa con altre emozioni, come la gioia e la rabbia, con cui condivide una serie di caratteri prosodici (Cahn, 1990). Questi studi concordano nell'associare la sorpresa ad alti livelli frequenziali, una certa variabilità di F0 e un'intensità maggiore rispetto al parlato neutrale. Infine, si rileva che mentre nello svedese la sorpresa è interessata da allungamento temporale, nello spagnolo si attesta una diminuzione della durata dei gruppi fonici. Relativamente all'italiano, le ricerca sulla comunicazione delle emozioni è in crescita, tuttavia i contributi che considerano l'emozione della sorpresa sono ancora pochi. Alcune informazioni si ricavano dai lavori incentrati sulla produzione e sulla percezione delle emozioni in italiano L2. In Maffia *et al*., 2014, l'analisi acustica ha confermato che, in quanto emozione ad alta attivazione, la sorpresa innesca un innalzamento dei valori di F0 e un'estensione del *range* tonale. Lo studio di De Marco & Paone (2014), dedicato alla codifica e alla decodifica delle sei emozioni primarie in apprendenti italiano L2, ha confermato che la sorpresa presenta le stesse caratteristiche prosodiche delle altre emozioni ad alta attivazione; per quanto concerne la percezione, lo studio ha rivelato che la sorpresa è identificata correttamente nella maggior parte dei casi (si rilevano, però, difficoltà più o meno significative a seconda della lingua nativa degli uditori).

Alla luce degli studi effettuati sorge un dubbio: la sorpresa, rispondendo a stimoli di natura diversa, può avere una connotazione positiva o negativa, a seconda che l'evento sia considerato vantaggioso o dannoso per il parlante, questa emozione può essere associata a caratteristiche psicologiche distinte ed espressa tramite mezzi diversi in relazione all'evento scatenante e al modo in cui il soggetto lo valuta. Tuttavia, finora, gli studi incentrati sulla realizzazione acustica del senso di sorpresa sono manchevoli di una distinzione fra le diverse connotazioni, positiva e negativa, di questa emozione. La presente ricerca intende, dunque, indagare la produzione e la percezione della sorpresa, specificamente ci si chiede: 1) Quali sono i correlati acustici della sorpresa con riferimento all'italiano di Bari? 2) La sorpresa positiva differisce da quella negativa in termini prosodici? La novità della ricerca consiste in primis nell'osservazione degli indici fonetici che in produzione differenziano la sorpresa positiva da quella negativa e in secondo luogo nello svolgimento di una verifica percettiva volta non alla valutazione del riconoscimento dell'emozione convogliata ma all'identificazione degli aspetti prosodici coinvolti nella decodifica del grado di sorpresa. La tipologia frasale scelta è l'esclamativa, che per definizione esprime la sorpresa del parlante rispetto alla realizzazione di un evento inaspettato: il significato della modalità esclamativa risiede proprio nella comunicazione di questo approccio del locutore alla realtà dei fatti. I materiali dello studio sperimentale constano di 20 sceneggiature ciascuna formata da un contesto in situazione, che dovrebbe suscitare sorpresa nel parlante, e una frase target (es. *Luca è arrivato!*). Di queste, 10 esprimono sorpresa positiva, poiché inserite in un contesto in cui è descritto un evento gradito, e 10 comunicano sorpresa negativa, poiché precedute dalla descrizione di un fatto spiacevole. Come controllo, è stato predisposto un campione di 20 frasi non connotate emotivamente, di tipologia frasale assertiva, sintatticamente e morfologicamente identiche alle frasi esclamative target. La sessione di registrazione ha coinvolto 10 giovani baresi (5 M e 5 F) invitati a leggere le sceneggiature a voce alta, nel modo più naturale possibile. I 400 stimoli ottenuti sono stati sottoposti a un'analisi acustica vertente sull'estrazione dei valori dei seguenti parametri, tramite il software *Praat* (Boersma & Weenink, 2016): valore medio (*F*0x), minimo (*F*0min) e massimo (*F*0max) della frequenza fondamentale (Hz), escursione melodica convertita in semitoni (ST), valore frequenziale dell'Onset e dell'Offset dell'enunciato (Hz), valore dell'intensità media (dB), durata totale dell'intero enunciato (ms), durata dell'ultima vocale tonica (ms), velocità di eloquio (sill/sec). Dall'analisi statistica effettuata tramite *Paired T Test*, è stato rilevato che la sorpresa è caratterizzata da valori di F0, intensità e durata significativamente maggiori rispetto al parlato neutrale, indipendentemente dalla connotazione emotiva; in particolare, i parametri più coinvolti sono l'estensione tonale, l'Onset, l'intensità media e la durata dell'ultima vocale tonica. Successivamente, il campione è stato organizzato in due gruppi in base alla connotazione affettiva della sorpresa espressa. I dati derivanti dal confronto statistico fra i due gruppi effettuato tramite *Paired T Test* (sorpresa positiva vs sorpresa negativa)

[1] Ekman (2003) ha messo in luce che soltanto sei emozioni sono riconosciute a livello universale. Tuttavia, a seconda dell'approccio adottato, il numero delle emozioni ritenute fondamentali può variare.

[2] Tuttavia, studi più recenti propendono per un approccio multidimensionale che individua oltre a quella vertente sul grado di attivazione, altre due dimensioni, ossia valutazione e potere (cf. Scherer, 2000, 2003).

hanno mostrato le frasi che esprimono sorpresa positiva presentano valori di *pitch range*, Onset e intensità significativamente più alti rispetto a quelle che comunicano sorpresa negativa. All'analisi acustica è seguita una fase sperimentale improntata alla percezione della sorpresa: ci si è chiesti se i parametri della frequenza e della durata siano coinvolti nella stessa misura nella decodifica dell'emozione sorpresa. Il test non verterà quindi sull'identificazione dell'emozione, ma sulla valutazione del grado di sorpresa espresso. Il disegno sperimentale ha previsto la manipolazione degli stimoli tramite sintesi vocale delle frasi, mediante il sistema PSOLA disponibile in *Praat*. Un sottocorpus di quattro frasi che esprimono sorpresa positiva è stato sottoposto a una manipolazione che ha comportato: a) l'accorciamento temporale della vocale nucleare in due step da 30 ms; b) l'abbassamento frequenziale dei primi 30 ms del contorno (Onset) in due step da 4 ST. I valori sono stati scelti facendo riferimento alle differenze medie riscontrate fra le esclamative e il campione di controllo; la manipolazione dei parametri è stata effettuata prima singolarmente e poi in maniera congiunta. Un test di 38 stimoli (4 originali, 24 modificati, 10 filler) è stato sottoposto per via telematica a 40 giovani uditori di provenienza barese. I partecipanti sono stati invitati ad ascoltare le frasi, presentate in ordine random, e a rispondere alla domanda "Secondo te l'enunciato esprime sorpresa?" tramite una scala Likert a 7 punti i cui estremi erano associati a "Per nulla" e "Moltissimo". I risultati, sottoposti ad analisi statistica tramite *Linear Mixed Model*, hanno dimostrato che il parametro dell'Onset e quello della durata sono particolarmente coinvolti nella trasmissione del senso di sorpresa, tuttavia, la manipolazione del primo sembra influenzare maggiormente gli uditori nell'elaborazione del giudizio.

Il presente studio si inserisce in un ambito scivoloso e complesso, quello della comunicazione delle emozioni, focalizzando l'attenzione su quella finora meno indagata, la sorpresa. I dati raccolti sottolineano l'importanza del ruolo svolto dalla *F*0x e dalla durata nella codifica e decodifica del senso di sorpresa, ma soprattutto fornisce i primi dati relativi alla distinzione fra prosodia della sorpresa gradita e sgradita, aspetto complesso, se si fa riferimento a un'emozione dai confini labili che spesso confina in altre emozioni.

Bibliografia

Abelin, Å. & Allwood, J. (2000), Cross linguistic interpretation of emotional prosody, *ITRW workshop on Speech & Emotion*, Newcastle, UK, sept. 2000, pp. 110-113.

Cahn, J.E. (1990), The generation of affect in synthesized speech, *Journal of the American Voice I/O Society*, 8, pp.1-19.

De Marco, A. & Paone, E. (2014), L'espressione e la percezione delle emozioni vocali in apprendenti di Italiano L2 uno studio cross-linguistico, *Educazione linguistica. Language education*, 3, pp. 483-500.

Ekman, P. (2003). *Emotions Revealed. Understanding faces and feelings*. London: Weidenfeld and Nicolson.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., & Longhi, L., Validation of an Acoustical Modelling of Emotional Expression in Spanish using Speech Synthesis Techniques, *ISCA Workshop on Speech & Emotion, Northern Ireland, 2000*, p. 161-166.

Izard, C. E. (1994). Innate and Universal Facial Expression: Evidence From Developmental and Cross-Cultural Research, *Psychological Bulletin*, vol. 115 (2), pp. 288-299.

Johnstone, T. & Scherer, K. R. (2000), Vocal communication of emotions, in Lewis, M. & Haviland, J. (Eds.), The *handbook of emotions*, New York: Guilford, pp. 226-235.

Juslin, P. N., & Laukka, P. (2003), Communication of emotions in vocal expression and music performance: Different channels, same code?, *Psychological Bulletin*, 129 (5), pp. 770-814.

Maffia, M., Pellegrino, E. & Pettorino, M. (2014). Labelling expressive speech in L2 Italian: The role of prosody in auto-and external annotation, in Campbell, A.W., Gibbon, D. & Hirst, D. (Eds), *Proceedings of the 7th International Conference on Speech Prosody*, Berlin, Germany, 20-23 May 2014, Speech Prosody Special Interest Group (SProSIG), Urbana, Illinois, pp. 81-84.

Scherer, K.R. (1986), Vocal affect expression: A review and a model for future research, *Psychological Bulletin*, 99 (2), pp. 143-165.

Scherer, K.R. (1989), Vocal correlates of emotional arousal and affective disturbance, in Wagner, H. & Manstead, A. (Eds) *Handbook of Psychophysiology: Emotion and social behavior*, London: Wiley, pp. 165-197.

Scherer, K.R. (2000), Psychological models of emotion, in Borod, J. (Ed.), *The Neuropsychology of Emotion*, Oxford: Oxford University Press, pp. 137-162.

Scherer, K.R. (2003). Vocal communication of emotion: a review of research paradigms, *Speech communication*, 40 (1-2), pp. 227-256.

Tomkins, S. (1963), *Affect Imagery Consciousness: Volume II, The Negative Affects*. London: Tavistock.

Tomkins, S. (1962), *Affect Imagery Consciousness: Volume I, The Positive Affects*. London: Tavistock.

Between-speaker variability in dynamic formant characteristics in spontaneous speech

Lei He[1] and Willemijn Heeren[2]
[1] *Department of Computational Linguistics, University of Zürich*
[2] *Leiden University Centre for Linguistics, Leiden University*
lei.he@uzh.ch; w.f.l.heeren@hum.leidenuniv.nl

Introduction

The temporal characteristics of speech articulation have received relatively little attention in forensic phonetics, because directly characterizing speaker-specific articulatory movements is almost impossible; kinematic data of articulators are absent from case materials. However, forensic speech scientists may instead focus on acoustic properties in the speech signal that are – although not entirely – modulated by the articulatory movements. For example, Dellwo and colleagues measured speech rhythm in terms of the durational variability of various phonetic intervals (e.g., Dellwo et al. 2015, Leemann et al. 2014) or syllabic intensity variability (e.g., He and Dellwo 2014, 2016); McDougall (2006) approached formant trajectories using least-squares polynomial approximations; and He and Dellwo (2017) measured the dynamic characteristics of intensity contours. Their study found that measures based on the speeds of intensity decreases (i.e., negative intensity dynamics) explained approximately 70% of between-speaker variability, pointing to a possibility that the mouth-closing gestures may contain more speaker-specific information.

More recently, He et al. (2019) combined the ideas of both McDougall (2006) and He and Dellwo (2017) and measured the dynamic characteristics of the first formant (F1). They found that the speeds of F1 decreases (reflecting mouth closing movements) contained more speaker-specific information than speeds of F1 increases (reflecting mouth opening movements). Moreover, an advantage of using F1 over intensity is that F1 measures are less affected by varying distances to the microphone. This is particularly relevant in forensic scenarios; voice experts typically have no information about the mouth-to-transducer distance, and distance may vary, in an unknown way, in the course of a recording. Moreover, the result that measures of negative F1 dynamics explained more between-speaker variability than measures of positive F1 dynamics is highly congruent to He and Dellwo (2017) using intensity dynamics.

However, He et al. (2019) only focused on Zürich German read speech in laboratory settings. To evaluate the practical value of this method for forensic practices, the current research aimed to test whether the same results will be obtained using spontaneous speech, in different languages. Thus, we aimed to investigate the generalizability of the findings from He et al. (2019) to scenarios much closer to the ones found in forensic speaker comparisons.

Method

Corpora and speakers

Vocalic nuclei were manually annotated in Praat (Boersma and Weenink, 2017) in data from three corpora, in different languages. This was done using phonetic transcripts created through forced alignment of available orthographic transcripts:
– For English, telephone conversations from 14 speakers were annotated (DyVis corpus [Nolan 2011], task 2). Per speaker, between 26 and 40 sentences were included (M = 33);
– For Dutch, spontaneous face-to-face conversations from 16 gender-balanced speakers were included (Spoken Dutch Corpus http://lands.let.ru.nl/cgn/ehome.htm). Per speaker, between 25 and 43 sentences were included (M = 34);

– For Zürich German, the TEVOID (Dellwo et al. 2015) corpus was used, containing16 gender-balanced speakers. Per speaker, 16 spontaneous sentences were extracted from an interview with an experimenter.

Acoustic and statistical analysis

The trajectories of F1 of each syllable nucleus were extracted using Praat (Boersma & Weenink, 2017), and the F1 dynamics (F1[+] and F1[–]) were calculated following the procedure described in He et al. (2019). The distributional characteristics of F1[+] and F1[–] in each sentence were calculated in terms of the mean (mean_F1[+] and mean_F1[–]), the standard deviation (stdev_F1[+] and stdev_F1[–]) and pairwise variability index (pvi_F1[+] and pvi_F1[–]). Multinomial logistic regressions were used to test the amount of between-speaker variability each of these measures can explain. This procedure was repeated for each of the languages.

Data processing and analysis are currently under way. We will present and discuss the results at the conference.

Acknowledgements

This work is being supported by an IAFPA research grant and an NWO VIDI grant (276-75-010).

References

Boersma, P; Weenink, D (2017) "Praat: doing phonetics by computer," Version 6.0.28, downloaded from http://www.fon.hum.uva.nl/praat/.

Dellwo, V; Leemann, A; Kolly, M-J (2015) "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *Journal of the Acoustical Society of America* 137: 1513–1528.

He, L; Dellwo, V (2014) "Speaker idiosyncratic variability of intensity across syllables," in *Interspeech 2014* , Singapore, pp. 233–237.

He, L; Dellwo, V (2016) "The role of syllable intensity in between-speaker rhythmic variability," *International Journal of Speech, Language and the Law* 23: 243–273.

He, L; Dellwo, V (2017) "Between-speaker variability in temporal organizations of intensity contours," *Journal of the Acoustical Society of America* 141: EL488–EL494.

He, L; Zhang, Y; Dellwo, V (2019) "Between-speaker variability and temporal organization of the first formant" *Journal of the Acoustical Society of America* 145: EL209–EL214.

Leemann, A; Kolly, M-J; Dellwo, V (2014) "Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison," *Forensic Science International* 238: 59–67.

McDougall, K (2006) "Dynamic features of speech and the characterisation of speakers: Towards a new approach using formant frequencies," *International Journal of Speech, Language and the Law* 13: 89–126.

Nolan, F (2011) Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006–2007 [data collection], UK Data Service.

Using Phonetic Theory to Improve Automatic Speaker Recognition

Elliot Holmes (University of York, ejh621@york.ac.uk)

Applications of Automatic Speaker Recognition are now widespread, being used in forensic voice comparison cases around the world whilst global commercial organisations (such as banks) use it for customer verification purposes. Within the field, there is a growing focus on the role that phonetic theory can play in complementing Automatic Speaker Recognition, specifically for understanding and improving upon current 'black box' approaches which developers cannot interpret. Such approaches have become popular in many other fields due to their increased speed and performance over interpretable systems and remain popular today; however, they are also proving increasingly problematic. For example, Rudin (2019) observed a 'black box' system used in hospitals to assess patient risk that, when it failed, put lives at stake. As the system was uninterpretable, the problem could not be identified nor rectified; however, once replaced with an interpretable system (which performed as well as the 'black box' system originally did), hospital staff are now able to understand any errors produced and rectify them.

Automatic Speaker Recognition does not need to take a similar risk; interpretable, phonetic features can be used to recognise speakers. Zhu et al. (2009) found that pitch features can discriminate between speakers; Long et al. (2011) found that harmonics-to-noise ratio can, and Ali et al. (2006) found that formants and bandwidths can. Studies have also shown that integrating interpretable phonetic features into current 'black box' systems, in particular those that measure laryngeal voice quality, can improve performance (Hughes et al., 2019).

The paper first presents a new methodology for testing phonetic approaches to improving Automatic Speaker Recognition. Current 'black box' approaches do not tailor, analyse, or specify any phonetic features or approaches before conducting Automatic Speaker Recognition tasks. The proposed methodology, however, is novel in that it specifies and extracts measurements of 35 phonetic features and measures them across phonetic units (phonemes) taken from every speech signal in two corpora. Then, comparisons are made between the two corpora, the contents of which can be tailored to the investigation at hand: this might be a comparison of two individual speakers, two gender groups, two age groups, or two accent groups. These two data sets are compared on a feature-by-feature, phoneme-by-phoneme basis using statistical modelling, specifically Linear Mixed Models (LMMs), to identify which linguistic features in each phoneme distinguish the two data sets. LMMs are more robust, as they allow for random effects structures to be specified. The aim is, in essence, feature selection, whereby we assess and rank features according to how often they are able to separate pairs of speakers in our dataset.

Then, this paper will utilise this methodology for a short study to demonstrate its usefulness: it will distinguish 100 speakers using the 35 features to compare tokens of their production of /a/. Specifically, every speaker was individually compared to the other 99 speakers using LMMs to identify which features distinguish their production of /a/. Nolan et al.'s (2009) DyViS Corpus has been selected for this study because there are 100 participants, all of which are comparable along sociolinguistic axes (gender, age, accent) and recorded with the same microphone. All tokens of the /a/ phoneme are text-dependent, having been taken from their production of the same passage of text.

This study presents findings that move towards a linguistically-informed improvement to Automatic Speaker Recognition: the methodology is novel and interpretable and its results identify phonetic

features in specific phonemes that are salient in distinguishing and characterising speakers from one another. Importantly, this methodology can now be replicated using bigger databases to expand upon the present example sample or even to identify distinguishing features between groups of speakers based on age, gender, and accent. Furthermore, it could also be used on shorter utterances or on text-independent speech to identify which phonetic features are salient to contexts more attuned to real-world uses of Automatic Speaker Recognition systems.

Reference List

Ali, A., Bhatti, S., and Mian, M. S. (2006, April, 22-23). *Formants Based Analysis for Speech Recognition*. [Paper]. IEEE International Conference on Engineering of Intelligent Systems (ICEIS), Islamabad, Pakistan. https://ieeexplore.ieee.org/ .

Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., and Gully, A. J. (2019). *Forensic voice comparison using long-term acoustic measures of voice quality.* [Paper]. International Conference of Phonetic Sciences, Melbourne, Australia. https://vincehughes.files.wordpress.com/.

Long, Y., Yan, Z., Soong, F. K., Dai, L., Guo, W. (2011, May, 22-27). *Speaker characterization using spectral subband energy ratio based on Harmonic plus Noise Model.* [Paper]. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic. https://ieeexplore.ieee.org/.

Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Forensic Linguistics,* 16(1). https://www.researchgate.net/.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence,* 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x

Zhu, J., Sun, S., Liu, X., and Lei, B. (2009). *Pitch in Speaker Recognition.* [Paper]. Proceedings of the 2009 Ninth International Conference on Hybrid Intelligent Systems, Massachusetts, U.S.A. https://dl.acm.org/.

Formant variability in five Hungarian vowels with regard to speaker Discriminability

Anna Huszár[1], Valéria Krepsz[1], Alexandra Markó[2,3] and Tekla Etelka Gráczi[1,3]
[1]Hungarian Research Institute for Linguistics, [2]ELTE, Department of Applied Linguistics and Phonetics,
[3]MTA-ELTE Lendület Lingual Articulation Research Group

Introduction. Formant frequencies are often used in speaker identification [e.g. 1, 2, 3], and higher (F3, F5) formants were hypothesised and found to be more distinctive among speakers [e.g. 2, 3]. [4] tested three vowels' first five formants separated and in combination of 2 and 3 of them. When one formant was included in the test F4 or F5 gave the best classification rate depending on the vowel, when two or three formants were included the combination giving the best classification rate always included one or two of F2, F4 and F5 [4]. The test and train speech materials are often of the same speech style, as their combination may raise further questions, like the possible difference on the scale of hypo-hyperarticulation [5] and speech accommodation [e.g. 6]. The present study's main questions are how sentence repetition and reading aloud may differ with regard to vowel realisations and how speaker identification is affected by the speech style difference. Our question was how the vowel formants differ across the speech styles, and if the possible difference affects the across speech styles speaker classification rate. We hypothesised that reading aloud will result in more distinct formant values. We also hypothesised that varying the vowel and formant values used for LDA will impact the classification rates.

Methods. 30 male and 30 female speakers aged between 20 and 50 years were selected from the Hungarian "BEA" database that includes the recordings of various speech styles of 461 speakers [7]. The recordings were carried out in a sound attenuated room by an AT4040 microphone. Two speech styles were selected: (i) repeated and (ii) read sentences. The sentences were the same in the two tasks. In the first task, the interviewer reads these one-by-one and the speaker repeats them by ear. In the second task, the speaker reads the sentences aloud. The same 20-20 CVC-realisations of the 5 most frequent Hungarian vowel phonemes /ɒ a ɛ e i/ were analysed in the two tasks (the 60% of the V-realisaions appeared in opened syllables and 50% of them in the first syllable that is the position of the word stress). Altogether 200 vowel realisations per speaker. The first five formants were measured at the mid 50% of the vowels' duration automatically by a Praat script [8]. The basic settings were 5 formants in 5500 Hz for women, and 5 formants in 5000 Hz for men. However, in the cases where more than 2 false values appear in a vowel quality in the same formant the settings were changed to the most fitting found by manual measurement probes. The possible differences between the speech styles were tested by linear mixed models [9, 10], the speaker discriminability was tested by LDA [11] in R [12]. The train set was one of the tasks, the test set the other task. All 5 vowels, one-one vowel was selected for training and testing in each model.

Results. The interaction of the task and the vowel was significant in the case of F1, F2 and F3 in women with showing smaller difference between the vowels, thus centralisation, hypoarticulation in repeated sentences than in reading. F4 and F5 were significantly higher in repeated sentences in the women's speech. F2, and F3 were significantly higher in the repeated sentences in men's pronunciation. F1, F4 and F5 did only show significant differences across vowels. Each formant was significantly different across vowels in both groups. The variability of F1 of /ɛ/ was considerably low among the speakers in both gender groups. The LDA were built with all 5 formants, and also with selecting 4, 3, 2 and 1 formant(s).

The classification rates of the LDA-models were higher for the vowels tested separately,as expected, but especially when including F1 and F2. The formant combination resulting in the best classification rate was vowel dependent, and was higher in male speakers in general. F5 did rarely contribute to the highest classification rates in contradiction to the results of [4].

Discussion & conclusions. The first three formants were less distinct among the vowels in women's pronunciation than in men's speech that may mean that women tend to hypoarticulate and centralise their vowels in repeated speech. The higher classification rates in general found testing men's vowel production may result from this difference. The higher classification rates in models including only one out of the analysed vowels evidently arises from the of the formant values' dependence on vowels. The lower classification rates for F5 than for lower vowels may appear due to the less accurate measurement and higher number of missing values due to automatic measurements. The results show in general that the speech style is an important factor in speaker identification.

References

[1] Jessen, M. 2008. Forensic Phonetics. Lang Linguist Compass 2(4), 671–711.

[2] Jessen, M. 1997. Speaker-specific information in voice quality parameters. Forensic Linguistics 4(1), 84–103.

[3] Nolan, F. 1983. The phonetic bases of speaker recognition. Cambridge: Cambridge University Press.

[4] Cao, H., Dellwo, V. 2019. The role of the first five formants in three vowels of mandarin for forensic voice analysis. In: International Congress of Phonetic Sciences, Melbourne, 5 August 2019 - 9 August 2019, 617-621.

[5] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the h&h theory. In: Hardcastle, W. J., Marchal, A. (eds.). Speech Production and Speech Modelling. 403-439.

[6] Eskénazi, M. 1993. Trends in speaking style research. Keynote speech. In: Proceedings of Eurospeech'93, Berlin. 501–509. http://www.cs.cmu.edu/~max/#_-_Publications.

[7] Neuberger, T., Gyarmathy, D., Gráczi, T. E., Horváth, V., Gósy, M., Beke, A. 2014 Development of a large spontaneous speech database of agglutinative Hungarian language. In: 17th International Conference, TSD 2014, September 8-12, 2014., Brno.

[8] Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer [Computer program]. Version 6.1.24, retrieved 29 September 2020 from http://www.praat.org/

[9] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48.

[10] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. Journal of Statistical Software 82(13), 1–26. http://doi.org/10.18637/jss.v082.i13.

[11] Venables W. N., Ripley, B. D. 2002. Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, http://www.stats.ox.ac.uk/pub/MASS4/.

[12] R Core Team 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

This research was supported by the National Research, Development and Innovation Office of Hungary, project No. FK-128814.

How robust are perceptual and acoustic observations of breathiness to mobile phone transmission?

Katharina Klug[1], Christin Kirchhübel[2], Paul Foulkes[1], Peter French[1, 3]
[1] Department of Language and Linguistic Science, University of York, UK
{kk667|paul.foulkes|peter.french}@york.ac.uk
[2] Soundscape Voice Evidence, Lancaster, UK
ck@soundscapevoice.com
[3] J P French Associates, York, UK

In forensic speaker comparison casework, analysts often rate a speaker's voice quality (VQ). In current casework practice in the UK as well as in Germany, observations about phonation features such as breathiness and creakiness are made using perceptual judgement. In an attempt to enhance assessments of VQ, the current study explores the possibility of using acoustic observations in tandem with auditory assessments, focussing on breathy voice.

Previous studies have identified spectral slope and additive noise measurements, such as H1*-H2*, H1*-A1*, HNR and CPP as acoustic correlates of the perception of breathiness (e.g. Klatt & Klatt 1990, Wayland & Jongman 2003, Garellek 2012, Hillenbrand et al. 1994). However, these results were based on high-quality studio recordings. The present study investigates how robust the acoustic measurements are, when dealing with degraded audio material typically found in forensic settings. Specifically, it explores the effect of GSM mobile phone transmission on the acoustic correlates of breathiness.

A set of 22 voices was created from six existing corpora of unscripted conversational speech from male speakers of British English (Gold et al. 2018, Haddican & Foulkes 2017, Kirchhübel 2013, Llamas et al. 2016-2019, Nolan et al. 2009, Wormald 2016). The voices were selected in order to reflect a natural mixture of non-pathological voices along the breathy/non-breathy continuum. The recordings were provided at a sampling rate of 44.1 kHz and 16 bit resolution and are referred to as 'microphone recordings' here. To generate forensically relevant degraded audio material, the same recordings were transmitted through a GSM mobile phone filter. These 'mobile recordings' are characterized by a typical frequency range of about 100-200 Hz to around 4000 Hz, depending on the specific source codec bit rate.

Two experiments were conducted, eliciting perceptual judgements. Four forensic speech analysts rated each voice in both conditions, microphone and mobile, in respect to presence/absence of breathiness as a dominant VQ feature. In total eight voices were consistently rated by all experts to lie at the extremes of the breathy/non-breathy continuum in both conditions: five dominantly breathy voices and three dominantly non-breathy voices. These voices qualified for acoustic analysis using VoiceSauce (Shue et al. 2011).

Findings suggest that auditory assessments of breathiness are robust to the telephone condition. In contrast, it appears that acoustic measurements are less robust. In the microphone recordings, H1*-H2*, H1*-A1* and CPP could be used to differentiate

between breathy and non-breathy voices. In the mobile condition, it appears that only CPP yielded significant differences between breathy and non-breathy voices while spectral slope measures did not. However, caution is needed when assessing precision and accuracy of the performance of the F0 and formant algorithms in the mobile filtered condition. Critical assessment and potential manual checking is needed.

Previous research has drawn attention to the fact that perceptual assessments of VQ demonstrate between- and within-rater variability (San Segundo et al. 2018). A combined auditory-acoustic approach would potentially reduce this variability. However, it remains a challenge to identify acoustic measurements that correlate with individual VQ settings and that can be accurately measured after mobile phone transmission.

References

Garellek, M. (2012). The timing and sequencing of coarticulated non-modal phonation in English and White Hmong. *Journal of Phonetics 31*: 152-161.

Gold, E., Ross, S. & Earnshaw, K. (2018). The 'West Y orkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. *Proc. 19th Interspeech* Hyderabad, pp. 2748-2752.

Haddican, W. & Foulkes, P. (2017). A comparative study of language change in Northern Englishes. [Data Collection]. Colchester, Essex: ESRC. URL: http://reshare.ukdataservice.ac.uk/851013/

Hillenbrand, J., Cleveland, R. A. & Erickson, R. L. (1994). Acoustic correlates of breathy voice quality. *J. Speech Lang. Hear. Res. 37*: 769-778.

Kirchhübel, C. (2013). The acoustic and temporal characteristics of deceptive speech, Doctoral dissertation, University of York.

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America 87*: 820-857.

Llamas, C., Watt, D. & French, J. P. (2016-19). The use and utility of localised speech forms in determining identity: forensic and sociophonetic perspectives. ESRC ES/M010883/1

Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law 16.1*: 31-58.

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V. & Kavanagh, C. (2018). The use of the vocal profile analysis for speaker characterization: Methodological proposals. *Journal of the International Phonetic Association 49.3:* 1-28.

Shue, Y.-L., Keating, P. & Vicenik, C., Yu, K. (2011). VoiceSauce: A program for voice analysis. *Proc. 17th ICPhS* Hong Kong, pp. 1846-1849.

Wayland, R. & Jongman, A. (2003). Acoustic correlates of breathy and clear vowels: the case of Khmer. *Journal of Phonetics 31.2:* 181-201.

Wormald, J. (2016). Regional variation in Panjabi- English, Doctoral dissertation, University of York.

A cross-linguistic study of between-speaker variability in intensity dynamics in L1 and L2 spontaneous speech

Carolina Lins Machado
Leiden University
carolina@machado.eti.br

Introduction

Earlier research investigated dynamic aspects of the amplitude envelope in Zürich German [1], [2]. Intensity dynamics were characterized as the temporal displacement of acoustic energy associated to articulatory mouth opening (positive) and closing (negative) gestures. Results showed that negative dynamics explained more between-speaker variability than positive dynamics. The current study examined positive and negative intensity dynamics, in spontaneous speech produced by Dutch speakers in both their native language (L1) and their second language English (L2).

The underlying mechanisms of speech production in the L1 and in the L2 are both similar and diverging [3], [4]. While the similarities are found in the mechanical apparatus used during speech production [5], the differences are related to the complex mechanisms taking place before articulation, which are believed to be affected differently by the different languages [6]. Along with language-specific constraints [7], the anatomical characteristics of a speaker are believed to influence how she/he uses her/his speech apparatus [8]. Therefore, a combination of individual characteristics and language constraints may affect the acoustic signal differently between speakers.

Thus far, intensity dynamics were studied solely in one language. Therefore, this study set out to determine (i) whether intensity dynamics vary between speakers of another L1, namely Dutch; (ii) whether this variability is also persistent in L2 productions by the same speakers; and (iii) whether language has an overall effect on intensity dynamics.

Method

Informal monologues in the L1 and L2 of 51 female speakers were selected from the LUCEA corpus [9]. The speech was manually annotated orthographically by two annotators and checked by a third annotator. Next, duration of analysis chunks and intensity were normalized by language. For duration, the data was chunked to obtain uninterrupted speech segments of ca. 1.5 seconds, see [10]. For intensity, its curve was linearly normalized following the method in [2], to maintain only information related to the curve's trajectory, which can be associated to speaker-specific articulatory gestures. Prior to the normalization of the amplitude envelope, the data was prepared according to the steps in [1] to obtain an object containing intensity point values in time. The detection of amplitude minima and maxima was done semi-automatically.

Measures of positive and negative intensity dynamics (mean [MEAN], standard deviation [STDEV] and sequential variability [PVI]) were calculated and extracted following [1], where positive dynamics refers to the rate of intensity increase from a trough to the next peak and negative dynamics to the slope in time between a peak and the next trough. Subsequently, statistical analyses of the extracted measures were carried out. A factor analysis (FA) evaluated the orthogonality of positive and negative dynamics, i.e. whether they encode different information. Then, a multinomial logistic regression (MLR) assessed how much between-speaker variability is explained by each type of dynamics. Next, a linear discriminant analysis (LDA) evaluated how well speakers are discriminated based on positive and negative measures of intensity dynamics. Finally, the effect of language on intensity dynamics' acoustics was investigated employing linear mixed-effects (LME) models.

Results and discussion

First, there was inter-speaker variability in intensity dynamics in both languages. However, the results in [1] for the FA and the MLR were not fully replicated in the present investigation. A possible explanation lies in the fact that this study used spontaneous speech. In the L1, the positive measure of sequential variability had a strong positive correlation with its negative counterpart (r = .80). This was

interpreted as a greater degree of gestural overlap between the start and end of syllables in spontaneous speech.

Similarly, the MLR results indicated that for both languages positive and negative dynamics seemed almost equally able to explain inter-speaker variability (48-52%). Across languages, negative dynamics explained a slightly larger quantity of inter-speaker variability, following the previously proposed [1] reduced prosodic control over the mouth closing movement. Results of the LDA displayed a low speaker classification rate in the L1 and L2; negative measures of mean were better classifiers for both languages (L1 = 4.8%; L2 = 4.4%, chance level: 1.9%). The results of the LME (Table 1) revealed an effect of language on all measures of intensity dynamics, suggesting differences on the rhythmic aspects of Dutch and English.

Table 1. *Condensed results of the LME models' estimates (standard errors) and 95% confidence intervals explaining the effect of language on positive [+] and negative [–] measures of intensity dynamics (mean [MEAN], standard deviation [STDEV] and sequential variability [PVI]).*

	β_0 (Intercept)		β_1 (language = English)	
	Est. (SE)	*[95% CI]*	*Est. (SE)*	*[95% CI]*
MEAN_v_I[–]	3.23 (.05)	[3.14, 3.32]	– .17 (.03)***	[–.23, –.12]
STDEV_v_I[–]	1.61 (.03)	[1.55, 1.66]	– .07 (.02)*	[–.11, –.02]
PVI_v_I[–]	6.28 (.06)	[6.16, 6.40]	– .72 (.05)***	[–.83, –.62]
MEAN_v_I[+]	4.27 (.07)	[4.13, 4.41]	– .13 (.04)***	[–.21, –.05]
STDEV_v_I[+]	2.19 (.04)	[2.11, 2.28]	– .08 (.03)**	[–.14, –.02]
PVI_v_I[+]	6.34 (.06)	[6.23, 6.45]	– .68 (.05)***	[–.78, –.58]

Note: Significance: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$, Bonferroni correction was applied.

Conclusion

Indexical characteristics in intensity dynamics were not restricted to the native language. Language had an effect on intensity dynamics' acoustics, indicating possible rhythmic differences between Dutch and English. Particularly, as in [1] and [2], negative intensity dynamics performed better than their positive counterparts in speaker discrimination methods and in explaining between-speaker variability.

Acknowledgment

This is the author's MA thesis, supervised by Dr. W.F.L. Heeren.

References

[1] He, L., & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *The Journal of the Acoustical Society of America, 141*(5), 488-494.

[2] He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *The Journal of the Acoustical Society of America*, *145*(3), 209-214.

[3] Levelt, W. J. M. (1999). Language production: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 83-122). Oxford, England: Oxford University Press.

[4] Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah: Lawrence Erlbaum Associates, Inc.

[5] Hixon, T. J., Weismer, G., & Hoit, J. D. (2020). *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. San Diego: Plural Publishing.

[6] Escudero, P. (2009). Linguistic Perception of ''similar'' L2 sounds. In P. Boersma, & S. Hamann (Eds.), *Phonology in perception* (pp. 151-190). Berlin: Mouton de Gruyter.

[7] Schwartz, G., & Kaźmierski, K. (2019). Vowel dynamics in the acquisition of L2 English – an acoustic study of L1 Polish learners. *Language Acquisition*, 1-28.

[8] Zuo, D., & Mok, P. P. K. (2015). Formant dynamics of bilingual identical twins. *Journal of Phonetics*, *52*, 1-12.

[9] Orr, R., & Quené, H. 2017. D-LUCEA: Curation of the UCU Accent Project Data. In Odijk, J., & van Hessen, A. (Eds.), *CLARIN in the Low Countries* (pp. 181-193). London: Ubiquity Press.

[10] Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628-639.

The first Italian Dysarthric Speech Database for improving daily living of severely dysarthric people

Marco Marini[1], Mauro Viganò[2], Massimo Corbo[2], Marina Zettin[3], Gloria Simoncini[3], Bruno Fattori[4], Clelia D'Anna[5], Massimiliano Donati[1], Luca Fanucci[1]

[1] Dip. di Ingegneria dell'Informazione, Università di Pisa, Pisa, IT

[2] Dip. di Scienze Neuroriabilitative, Casa di Cura Policlinico, Milano, IT

[3] Dip. di Psicologia, Università di Torino, Centro Puzzle, Torino, IT

[4] Dip. di Medicina Clinica e Sperimentale, Università di Pisa, IT

[5] Unità Otorinolaringoiatria audiologia e foniatria, AUOP, Pisa, IT

L'evoluzione dei sistemi di riconoscimento automatico del parlato (RAP) ha permesso l'utilizzo dei dispositivi intelligenti tramite comandi vocali. Questo traguardo è molto importante soprattutto per le persone che soffrono di patologie che ne limitano la mobilità, poiché gli permette di interagire con tali dispositivi senza doverli toccare o manipolare. Ma cosa accade quando la persona ha dei difetti di pronuncia o un linguaggio difficilmente interpretabile?

Le persone affette da disturbi dell'apparato fonatorio, come ad esempio la disartria, hanno difficoltà ad usare i sistemi RAP [1] [2] [3] poiché il loro modo di parlare è molto differente da quello di persone normodotate. Al fine di migliorare le performance di questi sistemi, la comunità scientifica negli anni ha realizzato diversi database di registrazioni di parlato disartrico. Infatti, gli algoritmi e modelli di intelligenza artificiale (IA) hanno bisogno di una grande quantità di dati per poter imparare e lavorare al meglio. Molti di questi database sono in lingua inglese, come ad esempio il Withaker [4] che contiene 19275 parole registrate da 6 persone disartriche dovuta a paralisi cerebrale. Oppure Nemours [5] e Universal Access speech (UASpeech) [6] e Torgo che contengono rispettivamente le registrazioni di singole parole e frasi molto corte di 11, 19 e 7 persone disartriche. Tuttavia, ad oggi non risulta esserci alcun database di parlato disartrico italiano e questo pone grandi limiti ai ricercatori italiani.

Per questo motivo, lo scopo della nostra ricerca è quello di realizzare il primo database di parlato disartrico italiano, utile per realizzare sistemi RAP usabili dai disartrici italiani. Per fare ciò, ci siamo rivolti a diversi enti e strutture ospedaliere per raccogliere più pazienti possibili con un ampio spettro di livelli di disartria. In particolare, le strutture con cui abbiamo collaborato sono 3: Azienda Ospedaliero-Universitaria Pisana (AOUP), il Centro Puzzle di Torino e la Casa di Cura Privata del Policlinico di Milano. Tale collaborazione ha due scopi principali: caratterizzazione del paziente che presta la propria voce (inserimento di dati anagrafici e clinici il più esaustivo possibile); registrare una quantità prefissata di parole per un certo numero di volte per il massimo numero di pazienti possibile.

Per la scelta delle parole da registrare, è stato deciso di affrontare uno scenario realistico e che preveda l'uso di parole singole e non troppo elaborate. Questo perché essendo il primo lavoro su parlato disartrico italiano, si è preferito affrontare una situazione un po' più semplice rispetto ad un riconoscimento automatico del parlato continuo che presenta molteplici criticità come le alterazioni prosodiche. Oltre a parole di uso domotico, si sono aggiunte altre parole al fine di coprire il più possibile tutti i fonemi italiani.

L'obbiettivo finale per i pazienti provenienti dalla AOUP è stato quello di registrare 45 parole per 3 volte (non consecutive) per un totale di 135 registrazioni a paziente, mentre per i pazienti provenienti dalle altre strutture, registrare 211 parole per 3 volte (non consecutive) per un totale di 633 registrazioni a paziente. Le 45 parole sono un sottoinsieme delle 211. La differenziazione tra la AOUP e le altre strutture è dovuta al tempo molto limitato che gli operatori sanitari e i medici hanno a disposizione con i loro pazienti e quindi non possono far registrare una grande quantità di dati. Dato che una persona disartrica solitamente è influenzata da altre patologie, un processo di registrazione potrebbe risultare molto stressante e faticoso, per questo motivo alcuni pazienti hanno interrotto la procedura di registrazione, quindi per loro alcune registrazioni sono mancanti. Le persone che sono riuscite a completare interamente la procedura di registrazione sono 21 (46% del totale), 8 persone (17% del totale) hanno registrato meno della metà delle registrazioni prestabilite, mentre 16 persone (35% del totale) ha completato più della metà delle registrazioni. Tutte le registrazioni sono state effettuate con lo stesso microfono a condensatore

[7], usando una codifica lineare in formato PCM a 16 bit e una frequenza di campionamento di 16kHz. Ogni parola registrata è salvata in un singolo file wave.
Per ogni paziente, prima di registrare qualsiasi parola, è stata compilata una cartella personale contente dati anagrafici (nome, cognome, genere e data di nascita) e clinici. Tale cartella è stata compilata da un medico competente. I dati clinici comprendono la patologia del paziente (è possibile scegliere la patologia in una lista di 9) ed altre informazioni aggiuntive come la data della diagnosi o la data dei primi sintomi. Inoltre, c'è la possibilità di compilare una scala di valutazione della patologia specifica qualora sia possibile. È altresì possibile specificare il tipo di disartria secondo la tassonomia di Duffy. La lista completa delle patologie selezionabili è la seguente: Sclerosi laterale amiotrofica, Parkinson e parkinsonismi, Corea di Huntington, Sclerosi Multipla, Distrofia Miotonica, Atassia, Ictus, Trauma Cranico, Neuropatia. Se il paziente non presenta alcuna di queste patologie, è possibile scegliere la voce "Altro" ed inserire delle note per spiegare la patologia specifica di quel paziente. Questi dati sono stati inseriti tramite un programma su computer provvisto di una semplice interfaccia grafica, appositamente realizzato dai ricercatori dell'Università di Pisa. Una volta profilato il paziente, il programma permette di iniziare a registrare la sua voce. L'intera procedura di registrazione è completamente assistita da un operatore che interagisce con il programma, il quale proporrà a schermo in modo automatico, la parola che il paziente deve ripetere ad alta voce. Se la registrazione non viene completata in modo corretto (rumore di fondo, il paziente ha pronunciato la parola sbagliata, ecc), l'operatore può far ripetere la registrazione al programma, oppure passare alla parola successiva nel caso in cui la registrazione viene effettuata correttamente. Il programma offre anche la possibilità di cambiare il contrasto della parola mostrata a schermo in modo da renderla più facile da leggere anche ai pazienti ipovedenti. Quando il paziente ha registrato tutte le parole almeno tre volte, il programma notifica il raggiungimento dell'obbiettivo e chiude la procedura. Se il paziente vuole continuare a registrare, può farlo liberamente e le parole registrate in più si aggiungono alle altre. Tutti i file di registrazione vengono salvati in locale nel computer dove è installato il programma e successivamente inviate a all'Università di Pisa direttamente dall'operatore.
In totale sono stati arruolati 45 volontari (25 maschi e 20 femmine) con 8 diverse patologie diagnosticate. Al 13,13% dei partecipanti è stata diagnosticata la sclerosi laterale amiotrofica, al 13,3% il Parkinson, al 2,2% la Corea di Huntington, al 15,6% un ictus, al 4,4% la sclerosi multipla, al 2,2% l'atassia e la distrofia miotonica, al 26,7% un trauma cranico, mentre al 20% altre patologie non presenti nell'elenco. In totale sono state registrate 13,72 ore di parlato. Sono già state effettuati degli esperimenti su questo database usando il toolkit Kaldi per la realizzazione di sistemi RAP. Tali risultati verranno presentati in fase di presentazione del progetto.
La nostra università ha finanziato un nuovo progetto di ricerca denominato DESIRE, composto da ingegneri, informatici, medici e umanisti con l'obiettivo di estendere e migliorare il lavoro già fatto, registrando non solo singole parole ma anche intere frasi formate da poche parole.

Bibliografia

[1] F. a. C. F. a. D. R. L. Ballati, «Assessing virtual assistant capabilities with italian dysarthric speech,» in *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 2018, pp. 93--101.

[2] D. a. M. G. a. M. M. a. F. L. Mulfari, «Towards a Deep Learning Based ASR System for Users with Dysarthria,» in *International Conference on Computers Helping People with Special Needs*, Springer, 2018, pp. 554--557.

[3] D. a. M. G. a. F. L. Mulfari, «Machine Learning in Assistive Technology: a Solution for People with Dysarthria,» in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, 2018, pp. 308--309.

[4] J. a. L. M. a. F. L. a. R. P. Deller Jr, «The Whitaker database of dysarthric (cerebral palsy) speech,» *The Journal of the Acoustical Society of America,* vol. 93, n. 6, pp. 3516--3518, 1993.

[5] X. a. P. J. B. a. P. S. M. a. L. J. E. a. B. H. T. Menendez-Pidal, «The Nemours database of dysarthric speech,» in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996.

[6] H. a. H.-J. M. a. P. A. a. G. J. a. H. T. S. a. W. K. a. F. S. Kim, «Dysarthric speech database for universal access research,» in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[7] Samson, «GoMic portable usb condenser microphone,» [Online]. Available: http://www.samsontech.com/samson/products/microphones/usb-microphones/gomic.

Álvaro Molina-García

Acoustics and Perception do not match in Andalusian Spanish

Medieval Spanish underwent a complex process of phonological merger in the order of sibilant fricatives. In Andalusia, this process led to the merger of CASA-CAZA (*House-Hunting*), a minimal pair that, however, has remained distinguished until today in standard Spanish. In the last century, Andalusia is undergoing a recovery of the CASA-CAZA distinction by mimesis of the national standard (Moya Corral y Sosińsky 2015; Villena Ponsoda y Vida Castro 2015, 2017; Regan 2017).

In our research on the reversal of the CASA-CAZA *merger* in Andalusian Spanish, we have verified an acoustic-perceptual dissociation. We have recorded 70% of standard realizations (contrast between CASA-CAZA) in perceptual analysis, versus just 37% of acoustic distance.

In order to know the social distribution of the process and the conclusions it implies for the principles of general phonology and language changes, the realizations of 54 informants were analysed. Those informants were classified classified by age (18-34, 35-54, >54 years), sex (male and female) and educational level (mandatory, intermediate, university). 20 words from the lexical series CASA, POSO and 20 words from the lexical series CAZA, POZO were analysed in each interview, with a total of 2160 realizations. Intervocalic realizations were always selected —in word interior (CA**S**A) or in initial position by syntactic phonetics (LA **S**ALA), distributed in a balanced way in the initial, middle and final part of the interviews.

Two different analysis were carried out. For perceptual analysis, all tokens were auditory tagged. The researcher classified the realizations according to a more sibilant [s] or non sibilant perception [θ]. Realizations from the CA**S**A lexical series perceived as sibilant [s] were considered split and those perceived as [θ] were considered merged. From 1080 tokens from CASA, 764 (70%) were perceived as [s] and 316 (30%) as [θ], so that it could be sustain that split is quite expanded.

For the acoustical analysis, words were cut and *spectral moments* (standard deviation, skewness, curtosis, centre of gravity), *spectral peak*, *intensity*, *duration* and *zero-crossings rate* were measured with Praat (Boersma y Weenink 2017). Then, multiparametric Euclidean distance was calculated from the sum of those acoustic parameters. However, the average of the acoustic distance was only 37%.

This acoustic-perceptual dissociation (70% in perception, 37% in acoustic) was clear in the following example. A previously regression analysis (carried out in order to explain the social distribution of the euclidean acoustic distance) reflected that the CASA : CAZA distinction had a social meaning associated with educated urban youth from affluent neighbourhoods (all those were the social independent variables included in the regression model: parents education, standard orientation, age, neighborhood, educational leve, modernity, etc.). However, as seen in Figure 1, despite the fact that the majority of speakers conformed to the system, there were two subgroups of speakers that do not; that is, they shown an acoustic distance lower than that expected given their social conditions.

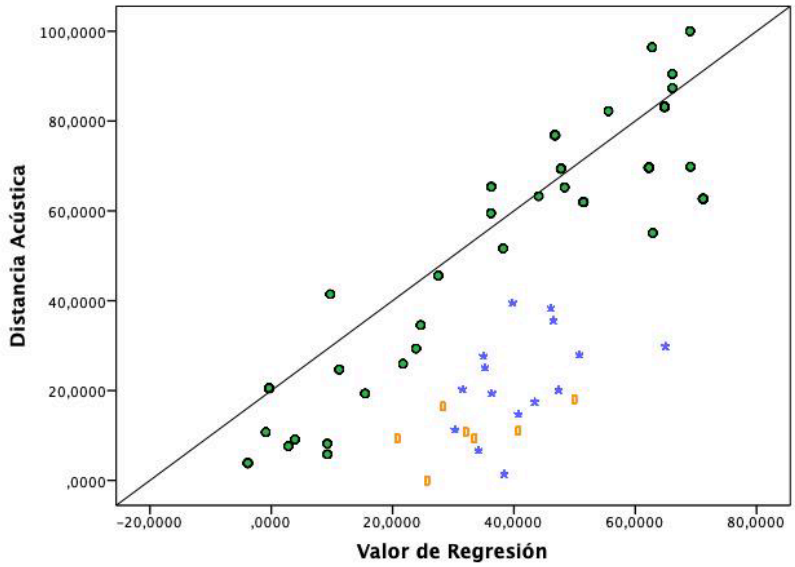


Figure 1
Dispersion of the speakers in the space defined by the mean acoustic distance and the effect of the social independent variables from the regression model

On the one hand, the **Independent Speakers** (7 orange rectangles), so called because their linguistic performance does not seem to depend on the social variables that make up the model, are made up of only 7 speakers. As shown in Table 1, these speakers not only show a low value of acoustic distance between CASA and CAZA, but also show quite low percentages of split performances in the auditory analysis. Although it is undeniable that it deserves an explanation, it should be simply the representation of the error that always exists within a multi-variant model. That is to say, since there is coherence between acoustics and perception, it could be considered that we are facing speakers with social conditions expected to present a relatively high percentage of standard realizations (contrast of CASA - CAZA), but that they do not achieve it, nor in acoustics or perception analysis.

However, the second subset of speakers, represented by blue asterisks (Graph 1), the **Dissociated Speakers** —so called because there is no correspondence between the results of the acoustic analysis and of the auditory analysis, is the most interesting case. This group, made up of 15 speakers, constitutes a case especially worthy of comment, since, unlike the previous subgroup, these speakers present very high percentages of split performances, but very low percentages of acoustic distance (Table 1); that is, an acoustic-perceptual dissociation is found between the results obtained in the auditory analysis and in the acoustic analysis.

The proposed explanation for this apparent contradiction is that a significant part of the speakers from the Malaga speech community would have advanced a great deal in the process of phonemic reallocation —i.e, they would have splitted the fricative continuum according to the lexical series CASA – CAZA because of a mimesis of standard Spanish, which means that individuals would tend to articulate more sibilant realizations for words from the CASA series, and less sibilant realizations for words from the CAZA series. However, since speakers have acquired since the last century the perceptual ability to differentiate allophones with small acoustic distance from each other, relatively close allophones aimed to represent the contrasting phonemes /s/ and /

θ/ are perceived as such by such native speakers. As speakers do not need to increase the allophonic distance to be perceived as *non-merger* speakers, only a small quantity of speakers, who potentially could do it, will not strive to increase this allophonic distance.

Independent Speakers			Dissociated Speakers		
Speaker	**Standard Realiz.**	**Acoustic Distance**	**Speaker**	**Standard Realiz.**	**Acoustic Distance**
5	0	9	**25**	70	1
22	5	25	**31**	85	12
52	10	9	**48**	90	14
13	35	18	**28**	95	18
21	45	16	**29**	95	10
3	50	11	**33**	95	19
20	50	11	**11**	100	0,38
			23	100	20
			26	100	28
			35	100	40
			36	100	30
			37	100	11
			39	100	35
			45	100	27
			53	100	20

Table 1
Groups of Independent and Dissociated Speakers. Comparison of the individual frequency of perceived standard realizations of the lexical series CASA : CAZA and the individual acoustic distance

A Comparative Analysis of Nigerian Linguist Native Speakers and Untrained Native Speakers Categorising Four Accents of Nigerian English

Umar Gombe Muhammad[1], Peter French[1,2], and Eleanor Chodroff[1]

[1]University of York, [2]JP French Associates

In the field of LAAP (Language Analysis in the Asylum Process), there has been a great deal of debate over who should undertake the task of inferring a speaker's country and region of socialisation based on their language and dialect: academic/professional linguists with detailed knowledge of the languages/varieties that may be at issue (see LNOG, 2004; Fraser, 2009; 2011), by naïve native speakers of those languages/varieties, or by a combination of both (see Cambier-Langeveld, 2010; 2012; Fraser, 2011; Foulkes, French and Wilson, 2019; Wilson, 2009). Opposing positions within this debate have largely been argued on principle alone, without support from empirical studies. Contributing to this debate, this presentation concerns research which forms part of a larger study designed to determine which of the following four methods performs most accurately and reliably in determining the first language (L1) of Nigerian speakers of English. It also seeks to establish a basis for combining selected methods.

(1) Review of speech samples by educated but linguistically naïve native speakers of the L1s concerned;

(2) Examinations of the samples by native speaker of the L1s who are also academic linguists holding staff positions in Nigerian universities;

(3) Examinations of the samples by general phoneticians and forensic phoneticians, with L1 British English, working either in UK universities or forensic speech science practices. (These participants are provided with priming material in the form of lists of phonetic features that distinguish the four Nigerian English varieties from one another);

(4) Analysis of the samples by an automatic accent recognition system (YACCDIST-Brown, 2016).

The material presented in the present abstract concerns only the relative performance of methods 1 and 2. It is anticipated that no method will perform with 100% accuracy. Assuming this, one might ask whether the errors of the highest performing methods are in complementary distribution such that, for example, the samples misidentified by naïve native speakers and those misidentified by UK phoneticians or by the automatic system do not (substantially) overlap. If that is indeed the case, one then has a principled and empirically-grounded basis for selecting which methods to combine in achieving optimum performance in LAAP casework.

Methods

Sixteen recordings made during a fieldwork visit to Nigeria of L1 of speakers of Hausa, Igbo, Kanuri and Yoruba speaking in English were selected for the experiment (categorisation task). These included four speakers from each of the four language groups; additionally, 2 foil recordings (Ghanaian and Guinean English speakers) were added. For Method 1, 80 linguistically naïve educated native speakers, mainly in the form of university students and administrative staff, were recruited in the universities and cities of Kano, Nsukka, Maiduguri and Ibadan.For Method 2, 25 academic linguists (with various specialisations in e.g.,phonetics/phonology, syntax, semantics and sociolinguistics of their Nigerian L1) were recruited from four Nigerian universities: Bayero University, Kano (Hausa linguist); University of Nigeria, Nsukka (Igbo linguists); University of Maiduguri (Kanuri linguists); University of Ibadan (Yoruba linguists). Using Qualtrics Survey Software, the two groups of participants were asked to listen to the recordings (which have equal duration of 30 seconds, comprising readings of part of a phonetically-balancedtest — The Rainbow Passage — and spontaneous speech narrating their life experiences) under equivalent conditions and assign each of the recordings to an L1 (Hausa, Igbo, Kanuri, Yoruba or non-Nigerian).

Results

Overall, native speakers from all 4 L1 backgrounds, irrespective of whether they were linguists, performed well above chance level; a mixed-effects logistic regression was used to assess the effect of L1 match and expertise (linguist or non-linguist) on language identification accuracy. L1 match was significant, indicating that listeners exposed to stimuli of their L1 were approximately 2.26 times as likely to be accurate than when exposed to stimuli of other L1s (β_{match} = 0.81, p<0.001). Native speaker linguists, however, were only slightly better than naïve native speakers on a numeric basis; the difference was not significant and applied to only 3 of the 4 L1 linguist groups (p > 0.05). Igbo linguists, who were the best performing group at identifying their L1 Igbo stimuli, were the poorest group at identifying other L1s (Hausa, Kanuri & Yoruba), while Kanuri linguists, who were the poorest at identifying their L1 Kanuri, were the best performing group at identifying other L1s (Hausa, Igbo & Yoruba). Hausa samples were easiest to identify by the linguists and non-linguists (well above chance level); Kanuri samples were the most challenging to identify by both groups (only at chance level). In a separate analysis, a set of salient L1 features were identified such as 'the' in Kanuri or 'h' in Yoruba. Samples which were richer in L1 cues (in terms of the number of pre-defined salient L1 features present in the recordings) were more correctly assigned to their L1s than others with fewer L1 features.

Discussion

Since the linguists were only slightly more accurate than the non-linguists in identifying the L1s, these findings offer empirical support for having educated native speakers involved in LAAP casework, even without linguistic training. Performance of both groups in identifying other Nigerian L1s (well above chance level) is facilitated by the listeners' familiarity with speakers of other L1s, owing to the movement of the participants in Nigeria and the use of English as lingua franca, as confirmed from them. Both groups were only around the chance level in identifying the two foils.The next stages of the research (testing the abilities of the UK phoneticians and an automatic accent recognition system-YACCDIST) are currently underway. In contrast to the previous methods, the UK phoneticians will also have the option of using auditory-perceptual or acoustic analysis help guide their categorisation of the speech samples.

References

Brown, G (2016). Exploring forensic accent recognition using the Y-ACCDIST system. *Proceedings of the 16th Speech Science and Technology Conference,* Sydney, Australia, 6-9 December 2016, 305-308.

Cambier-Langeveld, T. (2010). The role of linguists and native speakers in language analysis for the determination of speaker origin. *International Journal of Speech, Language and the Law, 17 (1), 67-93.*

Cambier-Langeveld, T. (2012). Clarification of the issues in language analysis: a rejoinder to Fraser and Verrips. *International Journal of Speech, Language and the Law, 19 (1), 95-108.*

Foulkes, P., French, P. and Wilson, K. (2019). LADO as forensic speaker profiling. In Patrick, P., Schmid, M. and Zwaan, K, (Eds). *Language Analysis for the Determination of Origin.*Cham: Springer, pp. 91-116.

Fraser, H. (2009). The role of 'educated native speakers' in providing language analysisfor the determination of the origin of asylum seekers. *International Journal of Speech*, *Language and the Law, 16 (1), 113–138.*

Fraser, H. (2011). The role of linguists and native speakers in language analysis for the determination of speaker origin: a response to Tina Cambier-Langeveld. *International Journal of* Speech*, Language and the Law, 18* (1), 121–130. https://doi.org/10.1558/ijsll.v18i1.121.

Language and National Origin Group. (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *International Journal of Speech, Languageand the Law, 11*(2), 261–266.

Wilson, K. (2009) Language analysis for the determination of origin: native speakers vs. trained linguists. MSc dissertation, University of York.

Dynamics of short-term cross-dialectal accommodation. A study on Grison and Zurich German

Elisa Pellegrino[1,2], Volker Dellwo[2]

[1]*URPP Language and Space, University of Zurich, Zurich, Switzerland*

[2]*Phonetics and Speech Sciences, Dept. of Computational Linguistics, University of Zurich, Zurich, Switzerland*

Accommodation, or the tendency of interlocutors to mutually adapt their linguistic behaviour during interactions or after increased exposure to communication partners, is a pervasive phenomenon in speech communication. If accommodation happens frequently enough between speakers of different dialects or accents short-term accommodation is hypothesised to bring about language variation and change (Trudgill, 1986). A study on vowel convergence between Swiss German dialects have shown that Zurich German (henceforth ZH) speakers converge more to Grison German (henceforth GR) speakers than vice versa, especially in low vowels and in words which served as stimuli in the dialogue (Ruch, 2015). This means that an innovation would occur in ZH dialect and this innovation would involve firstly low vowels, the most acoustically distant vowels between the two dialects. Understanding whether patterns of vowel convergence would echo in other cross-dialectal acoustic differences is thus of fundamental importance for understanding the diffusion of linguistic innovation and dialectal levelling in German speaking Switzerland. Therefore, in this paper we examine whether:

- cross-dialectal segmental temporal differences related to (a) open syllable lengthening (henceforth OSL), (b) geminate/singleton realization of intervocalic sonorants (henceforth ISG), (c) (un)reduced realization of unstressed vowel in word final position (RedVow) are prone to convergence insomuch as vowel quality (Eckhardt, 1991; Fleischer & Schmid, 2006);
- speakers of GR and ZH converge in segmental temporal properties in the same direction as for vowel quality;
- factors like acoustic distance can account for patterns of cross-dialectal phonetic convergence (Babel, 2010). Of the three durational contrasts RedVow is the most acoustically distant feature between the two dialects. Conversely, ISG and OSL are not dialect-specific since GR also admits the realization available for ZH (Table 1).

To study cross-dialectal temporal accommodation, we used the same corpus employed to examine vowel convergence. It comprises 18 audio-recorded dialogues between ZH and GR speakers who perform a diapix task, and 18 pre- and 18 post-dialogue recordings (picture naming task and retelling a story based on a comic), these latter performed individually by ZH and GR participants. To understand whether segmental temporal features evoke cross-dialectal convergence, we extracted lexical items instantiating the three target durational contrasts from the pre- and post-dialogue recordings. In pre- and post- dialogue recordings, we calculated three ratio measures devised to capture the cross-dialectal segmental temporal differences: (1) OSL: ratio between stressed and unstressed vowel within the same word; (2) ISG: ratio between intervocalic sonorants in -CCe words and in -Ce words; (3) RedVow: the ratio between word-final ending and stressed vowel. After that we calculated the Euclidean distances within pair and speaker before and after the interaction. Then, we calculated the difference in distance within a pair (*ddpair*) and within a speaker (*ddspeaker*). We expect that if patterns of vowel convergence replicate for segmental temporal properties, Euclidean distance between pairs of GR and ZH decreases after the interaction, ZH speakers converge more to GR and especially for RedVow (*ddspeaker* values lower than 0).

Preliminary results based on picture naming task show different patterns of accommodation between vowel convergence and segmental temporal properties. The data show that: (a) there are no differences in Euclidean distance within pairs before and after the interaction (fig. 1); (b) GR and ZH speakers did not show any distinctive pattern in the direction of accommodation (fig. 2), and in either measure (fig. 3). To conclude, interpretations of accommodation based on phonetic distance or geographical distribution are not tenable for segmental temporal differences in the present data. Vowel quality characteristics, which are more prone to convergence than segmental temporal ones, may play a major role in diffusion of linguistic innovations and dialectal levelling. Socio-linguistic

factors – e. g. speakers' attitude toward their own and the other dialect, perceptual salience of examined dialectal features, dialect markedness - will be also brough into play in the interpretation of documented dynamic of short-term cross-dialectal accommodation.

Dialectal feature	Example with transl.	GR realization	ZH realization
ISG	Sonne 'sun'	nn ['sunnɐ] n [sunɐ]	n ['sunə]
OSL	Sohle 'sole'	V: ['so:lɐ] V ['solɐ]	V ['solə]
Red Vow	Suppe 'soup'	ɐ ['suppɐ]	ə ['suppə]

Table 1. Examples of items in GR and ZH for the three durational contrasts

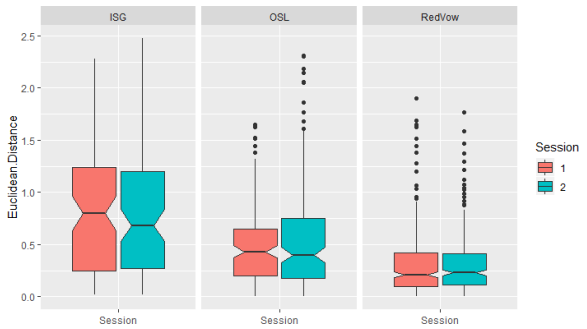


Figure 1 Euclidean distance within pairs across sessions (1 = before the interaction; 2= after interaction) for ISG (left), OSL (centre), RedVow (right).

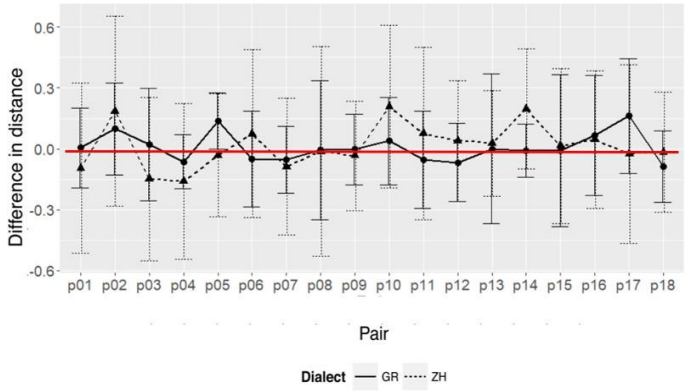


Figure 2: Difference in distance within speakers per pair and dialects.

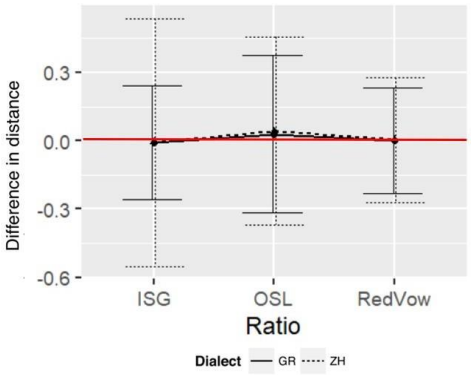


Figure 3: Difference in distance within speakers per ratio type and dialects.

References

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, 39(4), 437–56.

Eckhardt, O. (1991). Die Mundart der Stadt Chur. Zürich: Phonogrammarchiv der Universität 624 Zürich.

Ruch, H. (2015). Vowel convergence and divergence between two Swiss German dialects. *18th International Congress of Phonetic Sciences*, Glasgow UK.

Fleischer, J., & Schmid, S. (2006). Zurich German. *J. Int. Phon. Assoc*. 36, 243–253

Trudgill, P. (1986). *Dialects in Contact*. Oxford: Blackwell Publishing.

L2 speakers' individual differences in the acoustic properties of the front-high English vowels: The case of Ecuadorian speakers

Alejandra Pesantez Pesantez
University of Zurich, Switzerland
alejandra.pesantezpesantez@uzh.ch

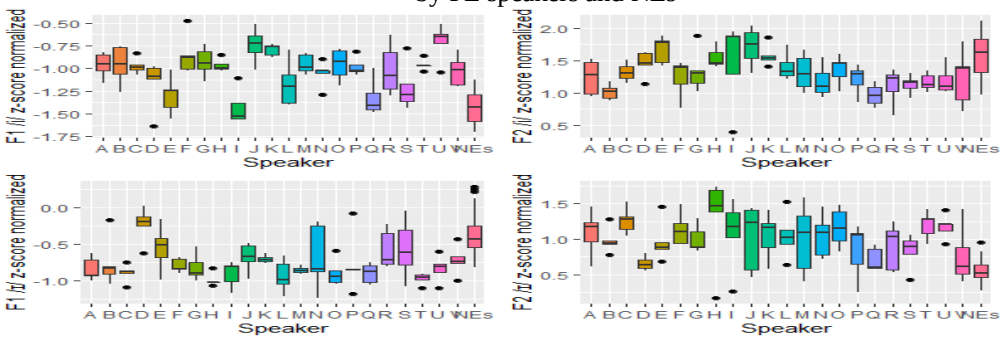
It is not surprising to find in a classroom setting second language (L2) speakers whose level of accuracy varies in the production of the target sounds. Most L2 speakers do not achieve near native-like pronunciation, especially if they learn the language during their adulthood [7]. Apart from the age of acquisition, some studies have demonstrated how learners' L1 backgrounds influence the accuracy of the L2 vowels. For example, L1 Spanish speakers of English fail to produce more accurately the English vowel sounds because they rely on vowel duration as acoustic cues to produce differences between the English tense and lax vowels, even though contrast on durational cues does not exist in the Spanish language [5, 12, 13]. Trying to explain which other factors impact the L2 acquisition, a large and growing body of literature has investigated the role of input [8, 10], the L2 language experience [1, 4, 9] the L2 aptitude [11, 13], and motivation [2]. Most of these studies have reported data on L2 acquisition in a naturalistic setting, indicating that as experience (numbers of years living in an English-speaking country) and input (exposure to native-speakers) increase, L2 English speakers can eventually learn to produce more accurately the sounds of the target language. At the same time, cross-sectional studies on group-level observations cannot describe how much individual variability exists between L2 speakers, especially when they learn English as a Foreign Language (FL), getting in most cases experience in the classroom with native and non-native teachers.

The aim of this study is to explore FL speakers' individual variability in the accuracy of the English vowel /i/-/ɪ/ produced by L1 Ecuadorian Spanish speakers and their ultimate attainment of the target vowel sounds acquired in a classroom setting. Two groups of speakers were recruited for this study. The native American English (NEs) group included 3 females and 2 males with a mean age of 24.4 years. The FL speakers group consisted of 22 L1 Ecuadorian Spanish monolingual speakers (16 female and 6 male adults) enrolled in the eighth level of the English Language Teaching program with a mean age of 25.9 and who had participated in the Phonetics and Phonology courses. 40 monosyllabic words containing the English vowel contrasts /i/-/ɪ/, /u/-/ʊ/, /ɛ/-/æ/, /ʌ/-/ɑ/ in a CVC and CVCC context were produced by the FL speakers. The same task was performed by the NEs. For this study, we only analyzed the tense and lax vowels /i/-/ɪ/. The recording session took place in the radio station of the university of Cuenca-Ecuador. A picture-naming task was used to elicit the target vowels and words were written in Spanish to avoid the effect of orthography in the production of the English segment phonemes [5, 8]. We used a Zoom H2n handy recorder at 44.1 kHz sampling, 16-bit quantization. To compare the spectral characteristics between each L2 speaker and NEs, we first manually annotated the vowel segments, using Praat software [3]. For individual vowels, we automatically extracted the mean of F1and F2 using a Praat script. The two formant values were Lobanov z normalized, and two mixed effect models were applied for each vowel and speaker.

A visual inspection of the data indicated significant individual variability in the accuracy of the English vowels /i/-/ɪ/, ranging from some speakers who produced F1 and F2 values truly separated from other FL speakers and some who followed the trend of the group [Fig.1]. The results of the two mixed effect models applied to each FL speaker and compared with NEs demonstrated that for the /i/ tokens 3 speakers got F1 values which were close to NEs criteria and 10 speakers produced F2 formant values within the same range as NEs did. For /ɪ/ tokens, 16 speakers got F1 mean values lower than NEs criteria. For the F2 mean values, most FL speakers did not reach NEs criteria, but two speakers

got F2 values similar to the model. The single most striking observation to emerge from the data was that no FL speaker could produce accurate F1 and F2 values for the two groups of vowels compared. These results demonstrate that not all the formants have the same level of difficulty to be produced by FL speakers, and speakers end up with different acoustic variability in terms of their ultimate attainment in a classroom setting.

Fig 1: Boxplot of the F1 and F2 formants values of the /i:/-/ɪ/ English vowels produced by FL speakers and NEs



Keywords: vowel accuracy, FL speaker, individual variability, classroom setting.

Acknowledgement

This project was funded partially by the University of Zurich-Graduate Campus, GRC Travel Grants.

References

[1] Baker, W., Trofimovich, P. 2006. Perceptual paths to accurate production of L2 vowels: The role of individual differences. *IRAL–International Review of Applied Linguistics in Language Teaching*, 44(3), 231-250.

[2] Basuki, Y. 2016. The Use of Peer-Assessment of Reading Aloud to Improve the English Department Students' motivation on Pronunciation Class of STKIP PGRI Trenggalek. *Jurnal Pendidikan Dewantara*, *2*(1).

[3] Boersma, Paul., Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version6.1.09, retrieved 26 January 2020 from http://www.praat.org/

[4] Colantoni, L., Steele, J., Escudero, P., Neyra, P. R. E. 2015. *Second language speech.* Cambridge University Press.

[5] Escudero, P., Boersma, P. 2004. Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 551-585.

[6] Escudero, P., Wanrooij, K. 2010. The effect of L1 orthography on non-native vowel perception. *Language and speech*, *53*(3), 343-365.

[7] Fabra, L. R., Romero, J. 2012. Native Catalan learners' perception and production of English vowels. *Journal of Phonetics, 40(3),* 491-508.

[8] Flege, J. E., & Fletcher, K. L. 1992. Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America*, *91*(1), 370-389.

[9] Flege, J. E., Bohn, O. S., Jang, S.1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of phonetics, 25*(4), 437-470.

[10] Flege, J. E., Wayland, R. 2019. The role of input in native Spanish Late learners' production and perception of English phonetic segments. *Journal of Second Language Studies, 2(1),* 1-44.

[11] Huensch, A., & Thompson, A. S. 2017. Contextualizing attitudes toward pronunciation: Foreign language learners in the United States. *Foreign language annals, 50(2),* 410-432.

[12] Kondaurova, M. V., Francis, A. L. 2008. The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*(6), 3959-3971.

[13] Morrison, G. S. 2008. Perception of synthetic vowels by monolingual Canadian-English, Mexican-Spanish, and Peninsular-Spanish listeners. *Canadian Acoustics*, *36*(4), 17-23

[14] Saito, K., Plonsky, L. 2019. Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652-708.

Variazione e *user engagement*. Un approfondimento sulla ludicizzazione dei protocolli d'inchiesta linguistica

Duccio Piccardi (Università degli Studi di Siena)
Fabio Ardolino (Università di Pisa)

I. Introduzione

Piattaforme, tecnologie e risorse informatiche hanno conosciuto nell'ultimo decennio una diffusione rapida ed esponenziale, sia nei campi lavorativi sia in quelli legati al tempo libero e all'intrattenimento. In linea con questa tendenza, sono molte le discipline che oggi si stanno occupando dei cosiddetti processi di ludicizzazione (ing., *gamification*): un termine che designa l'utilizzo di elementi strutturali tipici della progettazione videoludica (sistemi a punti, missioni, classifiche) in contesti tradizionalmente non ludici [1] al fine di stimolare l'utente al completamento di determinati compiti [2]. Mettendo a punto protocolli d'indagine ludicizzati, la ricerca scientifica tenta di stimolare la motivazione ed il coinvolgimento del partecipante al fine di migliorare la qualità dei dati raccolti; partecipanti più coinvolti dedicheranno al task specifico più attenzione, e saranno meno propensi ad abbandonare il compito prima di averlo completato.

La linguistica si sta recentemente affacciando ad una riflessione critica sulla ludicizzazione dei protocolli d'inchiesta. Un numero speciale di *Linguistic Vanguard* (https://linguistlist.org/issues/29/29-4100/), attualmente in preparazione, segue la realizzazione di studi ludicizzati su, fra le altre cose, morfosintassi [3], parlato emotivo [4], sociolinguistica e diacronia [5] e fonetica [6]. Con riferimento a quest'ultimo campo, la ludicizzazione dei protocolli rappresenta una potenziale risorsa nell'affrontare alcune delle più salienti problematiche sperimentali: su tutte, la difficoltà di elicitare parlato genuinamente spontaneo. Ciò detto, il lavoro da svolgere per comprendere appieno gli effetti degli approcci ludicizzati sulla raccolta di dati per la ricerca appare ancora lontano da un punto d'arrivo soddisfacente: in particolare, è spesso evidenziato il rischio derivante dalla sottovalutazione dei fattori di variazione latente insiti nella loro applicazione [7].

Partendo da questi presupposti, il nostro lavoro intende dimostrare che l'effetto della ludicizzazione dei protocolli sulle caratteristiche del parlato elicitato non è uniforme, ma dipendente dal coinvolgimento del partecipante nell'ambiente ludico dell'inchiesta. Prendendo le mosse dai metodi di indagine di una recente ricerca in ambito prosodico [8] e imperniata sull'uso del software *Minecraft: Education Edition* (MEE) [9], lo studio valuterà se la spontaneità del parlato elicitato, stimata a partire dalle realizzazioni di una variabile segmentale regionale (le occlusive sorde in posizione postvocalica a Firenze [10]), può essere predetta a partire dal coinvolgimento dell'utente, misurato per mezzo di uno strumento psicometrico apposito (UES, *User Engagement Scale* [11]).

II. Elicitare parlato tramite Minecraft: Education Edition

Minecraft [12] è un popolare videogioco per pc caratterizzato dalla possibilità, da parte del giocatore, di interagire in forme diverse con tutti gli elementi presenti nella mappa di gioco (i *blocchi*). La versione *Education Edition* (MEE) dell'opera nasce allo scopo di sfruttare la grande duttilità del mondo di gioco Minecraft a fini didattico- educativi. Le potenzialità di tale strumento, insieme alla relativa facilità d'uso, ha rapidamente portato MEE all'attenzione della ricerca scientifica [13], compresa la linguistica.

Nello studio [8], in particolare, gli autori hanno comparato le caratteristiche prosodiche di parole bersaglio elicitate tramite tre protocolli: un test psicoliguistico classico, lo stesso test ma con un interlocutore presente, e una sessione MEE. In quest'ultima, il partecipante era chiamato a risolvere alcune situazioni problematiche all'interno dell'ambiente di gioco attraverso l'interazione con gli elementi della mappa; alla soluzione di un quadro di gioco, l'utente veniva premiato con il passaggio al quadro successivo. Nel corso dell'esperienza, il soggetto era invitato a esporre a un altro giocatore, posto in modo da non vedere quanto accade sullo schermo del PC, le proprie ipotesi sulla soluzione dei rompicapi proposti, così come le azioni intraprese per completare il quadro di gioco: il parlato così prodotto costituiva dunque il dataset della condizione MEE. Gli autori descrivono il coinvolgimento del partecipante come uno degli aspetti chiave per spiegare le varianti prosodiche osservate: sebbene la presenza/assenza dell'interlocutore sia il principale discrimine per la modulazione dell'accentuazione prosodica, alcuni indici (intensità, estensione di F0) dipendono strettamente dal coinvolgimento nel task videoludico. In [14], gli stessi autori sottolineano come la ludicizzazione aumenti il coinvolgimento del parlante, portando quindi all'elicitazione di un parlato più spontaneo.

Nello studio che presentiamo in questa sede, adattiamo un protocollo MEE sul modello di [8] per cercare di approfondire due aspetti della ludicizzazione dei protocolli linguistici, ovvero: il coinvolgimento del singolo nell'esperienza di gioco, inteso come proprietà gradiente e non categorica, influisce sul parlato elicitato da un gruppo di partecipanti alla stessa condizione ludicizzata? E, inoltre, quali sono gli effetti della ludicizzazione a livello segmentale?

III. Valutare lo user engagement

[8] e [14] rappresentano il coinvolgimento come una caratteristica binaria, presente nell'inchiesta ludicizzata e assente in quella classica. In realtà, gli studi sull'interazione uomo-computer hanno evidenziato come il coinvolgimento sia un costrutto complesso, comprendente gli elementi cognitivi, emotivi e comportamentali coinvolti nella qualità e nella profondità di utilizzo di uno strumento informatico [15]. In questo senso, il coinvolgimento è da considerare come una proprietà gradiente, condizionata dalle caratteristiche del sistema, dal contesto d'interazione e dall'utente stesso; con riferimento a quest'ultimo aspetto, l'adozione di prospettive variazioniste mostra che anche le caratteristiche personali, come genere [16] ed età [17], possono avere un peso sul livello di coinvolgimento individuale.

Nell'approccio ludicizzato all'elicitazione di parlato, la sottovalutazione del coinvolgimento individuale rischia di confondere gli effetti legati alle fonti sociolinguistiche di variabilità. In linea con questa esigenza, nel nostro studio il coinvolgimento individuale è misurato per mezzo del questionario UES, uno strumento di autovalutazione rapido e pienamente validato, nonché recentemente aggiornato [11]. Il questionario consta di 12 quesiti a cui rispondere tramite scale Likert a cinque punti. Concettualizzando il coinvolgimento come una proprietà che si definisce nel momento stesso dell'interazione [15: 6], lo UES risulta particolarmente adatto ad essere utilizzato nel contesto di una singola sessione sperimentale. Nel nostro studio, i risultati dello UES vengono messi in relazione con la frequenza di realizzazione delle varianti delle occlusive sorde in posizione postvocalica nel parlato elicitato da partecipanti fiorentini durante l'esperimento.

IV. La ludicizzazione e il livello segmentale: tra iper- e ipoarticolazione

[8] trova nel parlato elicitato nelle condizioni sperimentali più interattive una maggiore discriminazione tra categorie prosodiche.

Per spiegare questo risultato, gli autori chiamano in causa due fenomenologie di natura diversa ma di esito prosodico analogo. Da una parte, il coinvolgimento porta a una maggiore spontaneità emotiva; dall'altra, i partecipanti estremizzano il carico funzionale di alcuni indici prosodici per farsi meglio comprendere dall'interlocutore. Nella condizione MEE, una corretta comunicazione è infatti essenziale per completare i livelli dell'esperimento.

Per contro, a livello segmentale, possiamo aspettarci che queste due fenomenologie concomitanti inducano effetti opposti sulla natura del parlato elicitato. Se da un lato la spontaneità può portare a un parlato più informale, dall'altra il contesto di gioco potrebbe invitare il partecipante a iperarticolare per aumentare le sue probabilità di successo nel compito (ad es. [18]). Nel parlato toscano, lingua e dialetto non sono entità distinte; caratteristiche dialettali emergono con minore o maggiore frequenza seguendo criteri stilistici [19]. L'indebolimento segmentale è uno degli elementi più caratteristici del parlato toscano [13]. La sua manifestazione più nota è la cosiddetta gorgia toscana, che interessa la realizzazione delle occlusive sorde [k t p] in posizione postvocalica. Nel nostro protocollo MEE, le parole bersaglio che i partecipanti saranno invogliati a elicitare durante la loro interazione con lo sperimentatore conterranno potenziali contesti di gorgia; per una migliore comparabilità con [8], non sono previsti filler. Seguendo quanto abbiamo detto finora, possiamo aspettarci due possibili esiti: a) più il partecipante è coinvolto nell'esperimento, più selezionerà varianti indebolite (non occlusive: fricative sorde ecc.) come conseguenza di una maggiore informalità; b) più il partecipante è coinvolto nell'esperimento, più selezionerà varianti pienamente occlusive nel tentativo di rendere il più efficiente possibile la sua comunicazione all'interno del gioco. In questo primo studio segmentale, le realizzazioni, analizzate sperimentalmente, saranno codificate in modo binario (*occlusive* vs. *non occlusive*). Al fine di disambiguare il ruolo del coinvolgimento sul tipo di risposta, ne valuteremo gli effetti su quelle categorie di partecipanti per le quali l'utilizzo di videogiochi non corrisponde a un contesto comunicativo abituale. Il successo dell'espediente sperimentale nel manipolare la formalità della realizzazione appare infatti dipendente anche dalla familiarità del partecipante con la specifica modalità impiegate nell'elicitazione [20].

Assumendo che una scarsa familiarità con l'attività proposta non pregiudichi necessariamente il coinvolgimento di un partecipante nella stessa [17], il lavoro ipotizza che l'impiego di un protocollo ludicizzato induca in modo più significativo forme di iperarticolazione segmentale in determinate tipologie di parlanti. Con questa premessa, l'analisi dei dati raccolti nell'inchiesta intende valutare il ruolo delle variabili *coinvolgimento* e *familiarità* (così come della loro interazione) nella produzione di parlato iperarticolato nel corso di elicitazioni ludicizzate (sul modello di [21]): a tale riguardo, è in ultima istanza esplorata l'ipotesi per cui nei parlanti meno familiari con l'intrattenimento videoludico esiste un effetto lineare del coinvolgimento sulla produzione delle varianti occlusive. L'analisi – condotta per mezzo di modelli generalizzati misti – è sviluppata in modo da includere anche variabili di controllo di natura linguistica (luogo di articolazione, *speech rate*) e sociale (sesso ed età del parlante).

[1] Deterding, S., Dixon, D., Khaled, R. & Nacke, L. (2011). From game design elements to gamefulness: Defining ''gamification''. In Lugmayr, A., Franissla, H., Safran, C. & Hammouda, I. (Eds.), *Proceedings of the 15th International Academic MindTrek Conference: Envisioning future media environments*. New York: ACM, 9-15.

[2] Sailer, M., Hense, J. U., Mandl, H. & Klevers, M. (2013). Psychological Perspectives on Motivation through Gamification. In *Interaction Design and Architecture(s) Journal*, 19, 28-37.

[3] Bos, J., Nissim, M. (2015). Uncovering noun-noun compound relations by gamification. In Megyesi, B. (ed.), *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*. Linköping: Linköping University Elektronic Press, 251-255.

[4] Yildirim, S., Narayanan, S. & Potamianos, P. (2011). Detecting emotional state of a child in a conversational computer game. In *Computer Speech & Language*, 25(1), 29-44.

[5] Leemann, A., Derungs, C. & Elspaß, S. (2019). Analyzing linguistic variation and change using gamification web apps: The case of German-speaking Europe. In *PLoSONE*, 14(12), 1-29.

[6] Leemann, A., Schmid, S., Studer-Joho, D. & Kolly, M. (2018). Regional Variation of /r/ in Swiss German Dialects. In *Proceedings of Interspeech 2018*, 2738-2742.

[7] Keush, F., Zhang, C. (2017). A review of issues in gamified surveys. In *Social Science Computer Review*, 35 (2), 147-166.

[8] Buxó-Lugo, A., Toscano, J. C. & Watson, D. G. (2018). Effects of Participant Engagement on Prosodic Prominence. In *Discourse Processes*, 55(3): 305-323.

[9] Koivisto, S., Levin, J. & Postari, A. (2012). *MinecraftEdu* [PC software]. Joensuu, Finland: Teacher Gaming.

[10] Marotta, G. (2001). Non solo spiranti. La 'gorgia toscana' nel parlato di Pisa. In *L'Italia Dialettale*, 62, 27-60.

[11] O'Brien, H.L., Cairns, P. & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. In *International Journal of Human-Computer Studies*, 112, 28-39.

[12] Persson, M., Bergenstern, J. (2011). *Minecraft* [PC software]. Stockholm: Mojang.

[13] Engelbrecht, J. A., Schiele, G. (2013). Koekepan: Minecraft as a research platform. In *12th Annual Workshop on Network and Systems Support for Games (NetGames)*, 1-3.

[14] Toscano, J.C., Buxó-Lugo, A., & Watson, D.G. (2015). Using game-based approaches to increase level of engagement in research and education. In Dikkers, S. (ed.), *Teachercraft*. Pittsburgh: ETC Press, 139-151.

[15] O'Brien, H.L. (2016). Theoretical perspectives on user engagement. In O'Brien, H.L., Cairns, P. (eds.), *Why Engagement Matters: Cross-Disciplinary Perspectives and Innovations on User Engagement with Digital Media*. Cham: Springer, 1-26.

[16] Vail, A.K., Boyer, K.E., Wiebe, E.N. & Lester, J. C. (2015). The Mars and Venus effect: the influence of user gender on the effectiveness of adaptive task support. In Ricci, F., Bontcheva, K., Conlan, O. & Lawless, S. (eds.), *User Modeling, Adaptation and Personalization*. Berlin/Heidelberg: Springer, 265-276.

[17] Fuschsberger, V., Sellner, W., Moser, C. & Tscheligi, M. (2012). Benefits and hurdles for older adults in intergenerational online interactions. In Miesenberger, K., Karshmer, A., Penaz, P. & Zagler, W. (eds.), Computers Helping People with Special Needs. Berlin/Heidelberg: Springer, 697-704.

[18] Schertz, J. (2013). Exaggeration of featural contrasts in clarifications of misheard speech in English. In *Journal of Phonetics*, 41(3-4), 249-263.

[19] Calamai, S. (2017). Tuscan between standard and vernacular: a sociophonetic perspective. In Cerruti, M., Crocco, C. & Marzo, S. (Eds.), *Towards a New Standard. Theoretical and Empirical Studies on the Restandardization of Italian*. Boston: De Gruyter, 213-24.

[20] Baugh, J. (2001). A dissection of style-shifting. In Eckert, P., Rickford, J. R. (eds.), *Style and sociolinguistic variation*. Cambridge: Cambridge University Press, 109-118.

[21] Duran, D., Lewandowski, N. (2020). Demonstration of a Serious Game for Spoken Language Experiments — GDX. In *Proceedings of the LREC 2020 Workshop Games and Natural Language Processing*, 68-78.

First indications for speaker individuality and speech intelligibility in state-of-the-art artificial voices

Claudia Roswandowitz[1,2,3], Thayabaran Kathiresan[2], Elisa Pellegrino[2], Volker Dellwo[2], Sascha Frühholz[1,3,4]

[1] Department of Psychology, University of Zurich, Zurich, Switzerland
[2] Phonetics and Speech Sciences, Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland
[4] Center for Integrative Human Physiology (ZIHP), University of Zurich, Switzerland

Introduction

In modern environment, situations in which humans encounter synthesized voices become more and more frequent. Individualist artificial voices have the potential to make human-computer interactions more natural and personalized. One application, among others, are individualist speech assistants for ALS or throat cancer patients who permanently lost their voices. However, to date there is little evidence on how vocal-identity and semantic speech information, both being fundamental for social interactions, are preserved in synthesized voices. First findings from speaker similarity rating studies suggest that indexical vocal information are poorly preserved in synthesized voices compared to the corresponding natural speaker identity (Lorenzo-Trueba et al., 2018). To address the open question, we used a modern voice-synthesis algorithm and tested speaker and speech recognition of artificial and natural speakers' voices.

Methods

To generate artificial voices, we used *sprocket* - an open-source voice conversion software using a Gaussian mixture model framework and a vocoder-free speech wave synthesis technique (Kobayashi & Toda, 2018). For sprocket, the voice conversion challenge in 2018 revealed second-best sound quality scores for same-speaker pairs and the sixth place for speaker similarity rating among 23 submitted conversion systems (Lorenzo-Trueba et al., 2018). Based on a parallel dataset (read speech material of ~30 minutes duration), Sprocket first learns the idiosyncratic acoustical features of a target and source speaker (i.e. MFCC, pitch) and then merges these features with the linguistic material of the source speaker. With this, we created high-quality artificial copies of four natural male target speakers (Standard German speakers, age range 19-34 years). As source speaker, we recorded speech material of one professional male speaker (Standard German speaker, 46 years).

Our experiment included a speaker familiarization task that was followed by the main experiment; the speaker speech matching task. 27 participants (mean age 25.58 years, 19 females) have first learned our four male speakers and after successful speaker familiarization (above 80%), 25 participants conducted the speaker speech matching task. This experiment comprised two task conditions: a speaker and a speech task and two voice conditions: natural and synthesized voices. In the speaker task, participants were asked to memorize the first target sound and to decide after each of the following test sentences (i.e. 12 test sentences per block, 20 blocks per condition) whether the sentence was spoken by the target speaker identity or another speaker. We presented 75 different 2-word declarative German sentences (e.g. "Er fällt.", "Er fehlt.") spoken by the previously familiarized speakers. The speech task follwed the same structure and the same sounds were presented, but this time participants matched the verbal content of a test sentence to the target sound, irrespective of who was speaking. We presented phonologically similar sentence to ensure comparable task difficulty between the speaker and speech task. The speaker speech task was conducted in an MRI environment and sounds were presented via active noise-cancelling headphones effectively reducing external MRI-induced sounds. To test for task and voice manipulation effects, we fitted linear mixed-effects models with fixed (task, voice condition) and random slopes (participants by voice condition) terms as implemented in the lme4 package (Bates et al., 2014) in the R environment (Team, 2019).

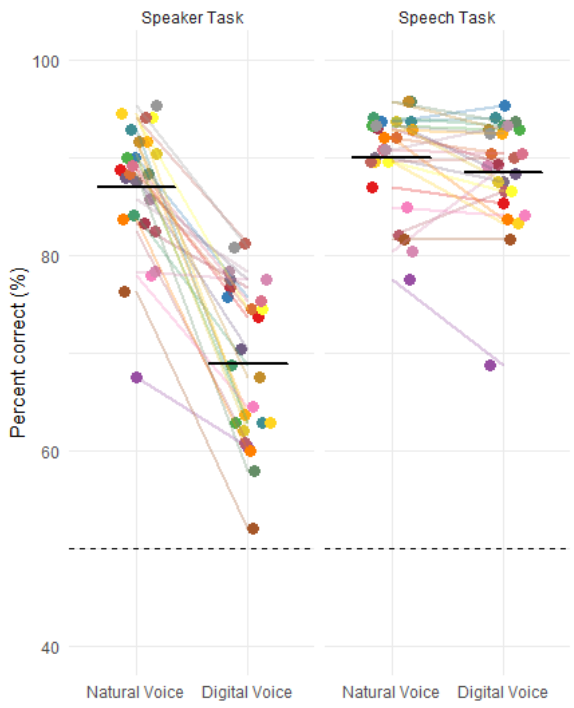


Figure 1. Individual results of the speaker and speech task. Solid black lines indicate mean percent correct for each task and voice manipulation. Dashed line indicate chance performance at 50%.

Results

In both tasks and voice manipulations, participants performed above chance level (i.e. 50%) indicating that speaker identity and speech recognition is possible for natural as well as synthesized voices (Figure 1). Next, we tested for an interaction between the speaker and speech task and the applied voice manipulation. As predicted, the interaction between task and voice manipulation (t = 9.04, p < 0.001) was significant, suggesting that the voice manipulation modulated the speaker and speech task differently. This difference is apparent in the speaker task with lower recognition performance for synthesized voices compared to the natural voices (t = 13.44, p < 0.001). Whereas in the speech task performance was not statistically different for the synthesized and natural voices (t = 1.03, p = 0.31). Overall, the model accounted for 82 % of the total variance in the data.

Conclusion

Our findings suggest that modern voice synthesis algorithms preserve socially relevant vocal attributes, especially semantic verbal information. However, we observed a marked reduction in identity recognition when listening to synthesized in contrast to natural voice identities. Our findings suggest that voice synthesis results in vocal versions of the natural speaker that are hardly accepted as natural variations of the corresponding natural speaker identity. Interestingly, speech recognition was largely unaffected by the voice manipulation. Our findings open new research avenues on social interactions in the digital age which may have important theoretical implications for current voice models but also technical application such as speech synthesize and automatic speaker recognition systems.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., & Haubo Bojesen Christensen, R. (2015). lme4: Linear mixed-effects models using Eigen and S4.

Kobayashi, K., & Toda, T. (2018). *sprocket: Open-Source Voice Conversion Software*. https://doi.org/10.21437/odyssey.2018-29

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. arXiv preprint arXiv:1804.04262.

Between-speaker variability in intensity slopes: The case of Thai

Yu Zhang[1], Lei He[1], Karnthida Kerdpol[2], Volker Dellwo[1]

[1]Phonetics & Speech Sciences Group, Institute of Computational Linguistics, University of Zurich, Andreasstrasse 15, CH–8050, Zurich, Switzerland
[2]Department of Linguistics, Naresuan University, Phitsanulok 65000, Thailand

Introduction

The major processes in speech production - glottal vibrations and articulatory movements - all contain speaker idiosyncratic information; such speaker idiosyncrasies leave traces in the speech signal and are thus measurable acoustically (see Dellwo et al., 2019 for a general review). Time-invariant properties of speech such as average fundamental frequencies (f0) and formant center frequencies have been extensively charted in speaker individuality research, and thus are generally accepted parameters of speaker identification decisions in forensic voice comparison and automatic speaker recognition (Singh and Murry 1978; Jessen 1997; Adami, Mihaescu, Reynolds and Godfrey 2003; Nolan and Grigoras 2005; Zhang, van de Weijer and Cui 2006; Lindh and Eriksson 2007; Morrison 2009; Leemann, Mixdorf, O'Reilly, Kolly and Dellwo 2014). However, speakers must evidently present some particular temporal features in the speech signals due to individual ways of moving their articulatory apparatus. He and his colleagues (2017, 2019) used positive and negative slopes to indicate both the speeds of intensity change and F1 change in a sentence and demonstrated that measures of negative slopes could explain more between-speaker variability in Zurich German (≈70%). Since intensity slopes and F1 slopes are both modulated by the opening-closing gestures of the mouth, these two acoustic measures are thus good estimates of mouth articulatory movements. Congruency in the findings indicates that the mouth-closing gestures during speech articulation may encode more speaker-specific information. Nevertheless, the previous findings (He and Dellwo, 2017; He et al., 2019) were obtained solely from native speakers of Zurich German. Investigating whether similar findings can be replicated from speakers of other languages is of great importance, especially when we wish to link the theoretical implications to forensic caseworks involving speakers of many languages. This paper thus followed the method with an improved statistical analysis and looked at how speaker differences are manifested in the temporal organizations of signal intensity curve using a Thai speech corpus.

Method

Thirteen native speakers of standard Thai (all female, aged between 20 to 22) were recorded reading the same set of 355 sentences (unidirectional microphone, sound-treated booth at Naresuan University, Phitsanulok/Thailand; 44.1 kHz, 16-bit). Following the same procedure in He and Dellwo (2017), positive and negative slopes (V[+] and V[–]) were calculated and then mean, variation coefficient (varco) and pairwise variability index (pvi) were used to characterize the distributions of positive slopes and negative slopes in a sentence. Z-score normalizations were performed for each particular measure to control sentence effect. The variance inflation factor (VIF) was computed for all six intensity slopes measures for diagnosing collinearity. The multinomial logistic regression (MLR) was fitted to quantify the amount of between-speaker variability explained by each of the intensity slope measures.

Results and discussion

The MLR results show that collectively measures of negative slopes explained 65.60% between-speaker variability (significantly higher than 50%, $\chi^2_{(1)} = 9.74$, $p < .01$). To directly compare the results with what He and Dellwo (2017) reported, we reanalyzed the Zurich German data and the MLR results showed that measures of negative slopes explained

68.44% (significantly higher than 50%, $\chi^2_{(1)} = 13.60$, $p < .001$) between-speaker variability in Zurich German, indicating that the results from the two languages were very similar.

The suprasegmental intensity or sonority fluctuations are one of the acoustic outcomes of mouth opening-closing movements, which organize speech into syllable-sized units constituting the rhythmic frames; this process has been argued to have an evolutionary advantage in speech comprehension (e.g., Chandrasekaran et al., 2009; MacNeilage, 1998; Morrill et al., 2012; Strauss and Schwartz, 2017). It is thus likely that the role of mouth opening-closing cycles is universal across languages, and the way speaker-specificity is encoded in the dynamic process and its acoustic outcomes are also similar in different languages. Needless to say, more languages should be investigated to testify our interpretation.

The findings have implications for research and applications where identity information in speech matters, such as forensic voice comparison (FVC) and automatic speaker recognition (ASR).

References

Adami, A., Mihaescu, R., Reynolds, R.D. and Godfrey, J.J. (2003). Modeling prosodic dynamics for speaker recognition, Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 788-91.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. and Ghazanfar, A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology 5, e1000436.

Dellwo, V., French, P., and He, L. (2019). "Voice biometrics for speaker recognition applications," in The Oxford Handbook of Voice Perception, edited by S. Frühholz and P. Belin (Oxford University Press, Oxford, UK), pp. 777--795.

He, L. and Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. The Journal of the Acoustical Society of America 141, EL488--494.

He, L., Zhang, Y., and Dellwo, V. (2019). "Between-speaker variability and temporal organization of the first formant," J. Acoust. Soc. Am. 145, EL209--EL214.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. Forensic Linguistics 4, 84-103.

Leemann, A., Mixdorff, H., O'Reilly, M., Kolly, M-J. and Dellwo, V. (2014). Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison. International Journal of Speech, Language and the Law 21(2), 343-370.

Lindh, J. and Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. Proceedings of INTERSPEECH 2007, 2025-2028.

MacNeilage, P. F. (1998). "The frame/content theory of evolution of speech production," Behav. Brain Sci. 21, 499--546.

Morrill, R. J., Paukner, A., Ferrari, P. F., and Ghazanfar, A. A. (2012) "Monkey lipsmacking develops like the human speech rhythm," Develop. Sci. 15, 557--568.

Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. The Journal of the Acoustical Society of America 125, 2387-2397.

Nolan, F. and Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. International Journal of Speech, Language and the Law 12, 385-411.

Singh, S. and Murry, T. (1978). Multidimensional classification of normal voice qualities. The Journal of Acoustical Society of America 64, 81-87.Strauss, A., and Schwartz, J.-L. (2017). "The syllable in the light of motor skills and neural oscillations," Lang. Cogn. Neurosci. 32, 562--569.

Zhang, C., van de Weijer, J. and Cui, J. (2006). Intra- and inter-speaker variations of formant pattern for lateral syllables in Standard Chinese. Forensic Science International 158(2-3), 117-124.

COARTICULATION AND VOT IN FOUR ITALIAN CHILDREN FROM 18 TO 48 MONTHS OF AGE

Zmarich C. [1,2], Bonifacio S., Busà M. G. [2], Colavolpe B. [2], Gaiotto M. [2], Olivucci F. [3]
[1]CNR-ISTC, Padova (I), [2]Università di Padova (I), [3]Università di Bologna (I)
claudio.zmarich@cnr.it

STATE OF THE ART. Over the past decades linguistic theories and theories of speech motor control have proceeded separately. This has been due to both theoretical and methodological reasons. From the theoretical point of view, there has been a pervasive influence of Generative Grammar, claiming a separation between Language and Performance, where Language systems are often regarded as developing independently of other cognitive and senso-motor systems, see Fodor, 1983). Methodologically, there has been a difficulty to trace neurobiological markers of the interaction between language and senso-motor systems, accompanied by an apparent lack of one-to-one correspondence between linguistic units and measures of executive behavior, see Smith (2010); Laganaro, (2018). Traditional explanations have emphasized the acquisition of phonology, lexicon and morphosyntax, but NOT of the motor processes related to speech production. The interconnection between linguistic factors and motor factors in the phonetic development is made complex by the continuous changes of the anatomo-physiological structures (for morphology, size and muscle innervation), of the neural substrate and of cognition (Callan et al., 2000). Yet the neural organization for sensorimotor and cognitive-linguistic aspects is highly interactive: for instance, the behavioral evidence shows a high degree of co-occurrence between cognitive-linguistic deficits and motor deficits (Goffman, 2015; McAllister-Byun, Tessier; 2016), and theoretical proposals linking together language planning, speech production and perception, and neurophysiological organization are currently available, covering also the developmental perspective (*DIVA* model, Guenther, 1995; Guenther & Vladusich, 2012; *State Feedback Model*, Parrell & Houde, 2019).

In order to study the acquisition of motor control in the early stages of language development, the empirical analyses preferably rely on acoustic data, as acoustic analysis has the advantage to quantify the phonetic continuum in the time-frequency domain, deriving information - by inference - on the underlying movements. More importantly, acoustic analysis is preferred because it is non-invasive, cheap and relatively simple to perform. Acoustic analysis is also useful for the study of coarticulation, which refers to the temporal overlapping of gestures belonging to neighboring phones (Hardcastle, Hewlitt, 2006; Farnetani, Recasens, 2010).

The present investigation aims to contribute some experimental data to two topics that have been understudied in Italian child language development, namely the acquisition of VOT (see Bortolini et al., 1995; Zmarich et al., 2009) and of anticipatory CV coarticulation (Petracco, Zmarich, 2006; Zmarich et al., 2009), though they have been thoroughly studied in other languages such as English, French or German (see Macken, Barton, 1980a; Eilers et al., 1984; Allen, 1985 for VOT; Sussman et al., 1999, for anticipatory coarticulation).

The best parameter for quantifying and classifying the voicing contrast is Voice Onset Time or VOT, which measures the time elapsed since the release of the consonantal occlusion to the beginning of the vibration of the vocal folds. Early in phonetic development, the voiced and unvoiced consonants tend to be realized as voiceless unaspirated, which allows the synchronization between glottal and supraglottal events; only after the acquisition of articulatory maneuvers children come to achieve the VOT categories that characterize the native language. In different languages the phonemic contrast between sonority categories corresponds to distinct temporal intervals along the VOT continuum (Abramson, Whalen, 2017; Cho, Whalen & Docherty, 2019).

While progressive coarticulation has been explained mainly as a result of articulatory inertia, anticipatory coarticulation has a cognitive base, because it must be planned in advance. According to the most influential hypothesis (Sussman et al., 2009), in the development of anticipatory coarticulation in a CV syllable, the child progressively narrows the domain of the articulatory organization from the syllable to the individual C and V gestures, with the consequence of decreasing coarticulation and increasing phonemic distinctiveness, but the process is not linear and depends on physiological constraints on the articulators. In fact, anticipatory coarticulation varies according to the articulatory place of the consonant. In the case of bilabial consonants, the articulation of the lips for the production of the consonant in CV syllable is not affected by the tongue dorsum during the production of the following vowel; this allows for maximum temporal overlap of the articulators (coarticulation as co-production). As for dental\alveolar consonants, the child must learn to differentiate and coordinate the apex (for the consonant) and the dorsum of the tongue (for the vowel), which are largely independent. As for velar consonants, the biomechanical constraints are maximum (both C and V are articulated with the tongue dorsum), and in this case the child must learn to mutually adapt the articulatory places for C and V (coarticulation as mutual adaptation).

AIMS OF THE PRESENT WORK. On the basis of both literature' results and our previous investigations, our working hypotheses are the following. VOT: Since the production of the initial voiced stop consonants require some mechanism external to the larynx in order to sustain an adequate transglottal pressure drop during the stop closure (as an active lowering of the glottis, Westbury, 1983), children will be more advanced in the acquisition of the appropriate VOT values for the voiceless than for the voiced consonants. Developmental studies on the acquisition of voicing in languages having voiced stops with negative VOT values (like Spanish (Macken & Barton, 1980b; Eilers et al., 1984)

and French (Allen, 1985)), show that two-year-old children have not acquired the VOT values for the initial voiced stops. Italian data from Zmarich et al. (2009) confirm the difficulty for some children even at the beginning of the fourth year of age. The acquisition criteria are the attainment of mean and standard deviation adult values, differing for the consonants place of articulation and the vocalic context (Bortolini et al., 1995). As for anticipatory coarticulation, we agree with Noiray et al. (2018): although coarticulation degree decreases with age, children will not organize consecutive articulatory gestures with a uniform organizational scheme (e.g. segmental or syllabic). Instead coarticulatory organization will be sensitive to the underlying articulatory properties of the combined segments (different lingual coarticulatory resistance and aggressiveness for consonants and vowels according to DAC model (Recasens, 1985), and subject to different articulatory constraints according to Sussman et al. (1999)). Criteria for acquisition are the attainment of mean adult values of coarticulation, differing on the basis of the consonants place of articulation.

METHOD. The data are part of a longitudinal corpus of ten children collected with the aims to study the typical speech development of Italian children. Four female subjects were recruited by one of the authors in Trieste (I), in two kindergarten centers from 2007 to 2009. The parents compiled the MacArthur CDI surveys for their children's lexical productions (Caselli et al., 2007) and filled out a questionnaire reporting information on normal psycho-physical development and monolingual language development (Italian). When they were 18-months-old, the children underwent audiometric screening (*Ling Six Sound Test*, Ling, 1976), to exclude the presence of hearing impairments. The children were recorded every three months from 18 to 48 months (11 sessions). The organization of the session was semi-structured (Schmitt, Meline, 1990), where the child interacted with the clinician in front of a set of toys. The objects were chosen based on the list of words compiled by the parent on the MacArthur CDI. In order to conduct a study on the development of VOT, in addition to saying "common" words, children were invited to repeat each of the 12 VOT test words at least three times. The test items were the following minimal pair pseudo-words, contrasting labial, dental and velar voiced and voiceless stops: *papa*, *baba*, *pipi*, *bibi*, *tata*, *dada*, *titi*, *didi*, *kaka*, *gaga*, *kiki*, *gigi*.

The acoustic files were annotated using *Praat*, through the use of *TextGrids*. All the productions with CV or CVC syllabic structures were selected. The borders marking the beginning of the consonant and the beginning of the vowel were also functional to the measurement of the VOT interval. Target and actual C and V individual segments were labeled. The syllable status as to lexical stress, the position in the word and the style of production (spontaneous or repeated) were also categorized. Finally, a number of exogenous events (like noise or uncertain transcription) or endogenous events, like a number of phonological processes, altering the syllable target, were also categorized.

The *Scriptgart.praat* developed by one of the authors was used to extract the VOT values (ms). For all the vowels present in the corpus, the script then proceeds to extract the values necessary for the calculation of the coarticulation, i.e. the F1 and F2 values in a number of points along the vowel, including the mid. Finally, for the consonant, the F2 values at the beginning of the formant transition are calculated. A number of other variables were obtained as a result of operations among the column of the final matrix, resulting in the duration values of segments and syllables (allowing to estimate the speech rate) and, most importantly, in the exclusion of syllables characterized by VOT values greater than +20 ms for the analysis of coarticulation. Only the syllables located at the beginning of the word that were preceded by silence (equal to or greater than 250 ms) were used for VOT analysis. As for anticipatory coarticulation, it is measured by means of *Locus Equations* (LE, see Lindblom, 1963; Krull, 1989). A LE describes a 1st order regression fit to a scatter of vowel steady-state frequency values predicting the onset of F2 transition values in CV sequences with a fixed C, of the form $F2_{cons} = k * F2_{vow} + b$. This measure provides an overall estimation of coarticulation, provided that LE slopes (indexed by *k* values) be calculated on CV sequences with vowel pooling and voiced plosives (Tabain, 2000). A nice characteristic of this method consists in an intrinsic normalization of *k* values, which could vary between 0 (no coarticulation at all) and 1 (maximal coarticulation), allowing the direct comparison of the production of different children at different ages those of adults.

RESULTS. The analysis of VOT from four children (Gaiotto, 2020; Colavolpe, 2020) revealed that differences exist in the chronology of the acquisition of voicing contrast according to the three places of consonant articulation. The variability of the VOT values within a single articulatory place could be an important index of the reorganization of the articulatory system. Most of the children made use of nasalization and fricativisation processes in order to start voicing in an attempt to reproduce voiced targets (like the French and Spanish children, respectively). The contrast of sonority does not assume important variations depending on stress or vowel context, but it seems to be sensitive to the style of production. As for coarticulation, the results confirm that the development of coarticulation can be better described by taking into account the biomechanical characteristics of articulatory gestures. The bilabials allow the maximum temporal overlap of the articulatory movements for which the development gradually proceeds towards a greater temporal synchronization, so that the back of the tongue can be found already in position for the vowel at the moment in which the lips open. Alveolars similarly require independent motor control maturation of two distinct portions of the same articulator (the tongue) to achieve a high degree of coarticulation. And finally velars show high levels of coarticulation from the beginnings, since the consonant and the vowel use the same articulator (the tongue dorsum).

Satellite Workshop

MICHAEL JESSEN *(Bundeskriminalamt, Wiesbaden, Germany, Language and Audio Division)*

Workshop on automatic and semiautomatic speaker recognition

In the first part of the workshop an overview will be given of the historical development of methods/systems that quantify comparisons between speakers in terms of a likelihood ratio. Some of the methods make use of input from acoustic phonetics (semiautomatic speaker recognition), others are based on cepstral coefficients, familiar from speech technology (automatic speaker recognition). The overview includes a demonstration of semiautomatic methods as well as automatic methods in their different historical stages (GMM-UBM, i-vector, x-vector) using the software VOCALISE. In the second part of the workshop participants have the opportunity to gain practical experience in the use of a speaker recognition system. Information regarding temporary licenses and access to speaker corpora will be provided directly to the workshop participants.

Round Table

Current trends and issues in forensic phonetics research

Moderated by:

PETER FRENCH
University of York, UK
https://www.york.ac.uk/language/people/academic-research/peter-french/

Participants:

VOLKER DELLWO
University of Zurich, Switzerland
https://www.cl.uzh.ch/de/people/team/phonetics/vdellw.html

HELEN FRASER
University of New England, Australia
https://helenfraser.com.au
https://forensictranscription.com.au

MICHAEL JESSEN
Bundeskriminalamt, Wiesbaden, Germany

KIRSTY MCDOUGALL
University of Cambridge, UK
https://www.mmll.cam.ac.uk/dr-kirsty-mcdougall

LUCIANO ROMITO
Università della Calabria, Italy
https://www.researchgate.net/profile/Luciano_Romito

Universität
Zürich[UZH]