Leibniz-Zentrum
Allgemeine Sprachwissenschaft

university of
groningen

Université
franco-allemande
Deutsch-Französische
Hochschule

Summer school 2022:

Coping with the complexity in speech production and perception

4-8 July 2022

Chorin, Germany

Organizers:
Susanne Fuchs
Anne Hermes
Leonardo Lancia
Martijn Wieling
Sarah Wesolek (local committee)

Book of Abstracts

Table of Content

Studying co-speech gesture as a dynamic sign(al)

Wim Pouw
wim.pouw@donders.ru.nl

The faculty of speech did not emerge in a vacuum. It co-evolved and always functions within a rich ecology of multimodal displays, signals, and signs (Deacon, 1998). One important aspect of the multimodal nature of speaking, is co-speech upper limb gesticulation, or gesture in short. Gestures are multidimensional. They beat with the rhythm of speaking. Gesture's rhythms signify our mood. And gestures pattern in complex ways to invoke deictic, iconic, and symbolic covariance relations, which if attuned to by a perceiver, leads to an expansion of the possible meanings that can be conveyed per unit of time.

Gestures are unwieldy in some ways, not only because of their time-varying dynamic nature, but also because of gestures' unspecific ways to iconically refer to absent state of affairs relative to the more conventionalized ways to refer to things in spoken or a signed language. As a consequence, the study of gesture has been very much dependent on skilled annotators to infer an intended referent from a gesture, or to treat gestures as non-continuous isolated events with an attributable mutually exclusive typology (beat vs. iconic gestures).

However, the field of gesture studies is becoming more heterogenous, increasingly drawing from other previously alien fields such as human movement, biomechanics, computer science. Firstly, there is a move away from mutually exclusive labels of gesture types, such that it is possible that a gesture has an iconic function while at the same time retain a beat-like quality (Shattuck-Hufnagel & Prieto, 2019). There is further more interest in the kinematic and physical study of gesture, supported by important advances in computer vision to track human multimodal behavior from audiovisual recordings (Pouw et al., 2020).

In this workshop I will overview some basic issues in gesture studies. Then I will motivate the value of approaching the study of co-speech gesture as a continuous phenomenon that can be an important additional level of analysis (Pouw, Dingemanse, et al., 2021; Pouw & Fuchs, 2021). At each step of the way, I will refer to open-source code and tutorials already online (e.g., Pouw & Trujillo, 2021), for participants to explore should it fit their concerns. This includes, code for running (light-weight) motion tracking on visual recordings (on your laptop) and to deidentify video data (Figure 1; (Owoyele et al., 2022)). Code tailored towards temporal and kinematic analyses of gesture. And code for visualizing all gestures that are produced in an intuitive way that can inform qualitative and support quantitative analyses (Pouw, de Wit, et al., 2021; Pouw & Dixon, 2019).

Figure 1: Example of how to reduce privacy risks when storing and communicating video data for research.

References

Deacon, T. W. (1998). *The symbolic species: The co-evolution of language and the brain*. W.W. Norton.

Owoyele, B., Trujillo, J., Melo, G. de, & Pouw, W. (2022). *Masked-piper: Masking personal identities in visual recordings while preserving multimodal information*. PsyArXiv. https://doi.org/10.31234/osf.io/bpt26

Pouw, W., de Wit, J., Bögels, S., Rasenberg, M., Milivojevic, B., & Ozyurek, A. (2021). Semantically Related Gestures Move Alike: Towards a Distributional Semantics of

Gesture Kinematics. In V. G. Duffy (Ed.), *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior* (pp. 269–287). Springer International Publishing. https://doi.org/10.1007/978-3-030-77817-0_20

Pouw, W., Dingemanse, M., Motamedi, Y., & Özyürek, A. (2021). A Systematic Investigation of Gesture Kinematics in Evolving Manual Languages in the Lab. *Cognitive Science*, *45*(7), e13014. https://doi.org/10.1111/cogs.13014

Pouw, W., & Dixon, J. A. (2019). Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. *Discourse Processes*, *57*(4), 301–319. https://doi.org/10.1080/0163853X.2019.1678967

Pouw, W., & Fuchs, S. (2021). *Origins of vocal-entangled gesture*. PsyArXiv. https://doi.org/10.31234/osf.io/egnar

Pouw, W., & Trujillo, J. (2021). *Envision Bootcamp 2021 Coding Modules*. https://wimpouw.github.io/EnvisionBootcamp2021/

Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020). The quantification of gesture–speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, *52*(2), 723–740. https://doi.org/10.3758/s13428-019-01271-9

Shattuck-Hufnagel, S., & Prieto, P. (2019). Dimensionalizing co-speech gestures. *Proceedings of the International Congress of Phonetic Sciences 2019*, 5.

Methods for the analysis of time series representing goal-oriented behaviour: A tutorial review

Leonardo Lancia, Laboratoire de Phonétique et Phonologie (UMR 7018), CNRS / Sorbonne Nouvelle
leonardo.lancia@sorbonne-nouvelle.fr

The number of quantities that are collected and analysed in the study of speech production and perception has grown considerably with respect to a few decades ago (when few acoustic parameters and even less articulatory quantities were systematically considered). On the one hand, this evolution is undeniably related to the possibilities offered by technological advances, thanks to which we can track articulator movements, target the behaviour of neural populations and collect a number of electrophysiological parameters with an unprecedented simplicity. On the other hand, such a rich portrait of speech activity is required by theoretical frameworks mapping the functioning of the sensorimotor system on that of complex dynamical systems relying on the task-dependent interactions between many partially independent quantities.

As a matter of fact, researchers often need to characterize the interactions between many heterogeneous quantities, potentially unfolding on different time scales, whose behaviour can change drastically due to these same interactions. And this of course is not an easy task. Several methods coming from other disciplines facing such kind of complexity (e.g., earth sciences, climatology, physiology, etc…) have been adopted in the analysis of human intentional behaviour and also in the analysis of speech behaviour. These methods are often conceived to unveil the features of relatively simple dynamical systems whose behaviour appears to be extremely different from that of the sensorimotor system during verbal interactions, or more generally during intentional behaviour. The difference is not only due to the potentially higher underlying complexity characterizing intentional behaviour in general but also, and potentially more importantly, to the fact that intentional behaviour (and especially speech) is accessible to symbolic descriptions capturing its most relevant features. The aim of my presentation will be that of introducing a unifying perspective on the analysis of simultaneously recorded time series based on the notions of cycles of activity, neighbouring states and recurrent states. This perspective provides a rationale to tailor to the analysis of verbal behaviour (and more generally of intentional behaviour) several methods designed to characterize the mechanisms underlying the behavior of continuous time series, as well as their mutual dependencies. We will adopt an empirical approach relying on rich supplemental material permitting to conduct all the illustrated computations with Matlab (and/or Python). However, our practical goal is not to introduce a toolbox of nonlinear methods for time series analysis that can be applied in an unproblematic fashion to any available dataset. On the contrary, we aim at focusing on the many limitations of the proposed tools and at providing the methodological awareness required to adapt these tools to different research objectives on the basis of domain-specific knowledge.

Approaches in Electromagnetic Articulography

Teja Rebernik, University of Groningen (t.rebernik@rug.nl)
Jidde Jacobi, Macquarie University & University of Groningen
Roel Jonkers, University of Groningen
Aude Noiray, Laboratoire Dynamique du Langage
Martijn Wieling, University of Groningen

Electromagnetic articulography (EMA) is a point-tracking technique that has been used for the study of speech production since the 1980s (e.g., Höhne et al., 1987). With sensors placed on the articulators (the tongue, the lips, the jaw), we get precise information on speech movements. EMA has thus been used for the study of different speech aspects, from the articulation of individual speech sounds and segments (amongst many e.g., Katz et al., 2017; Hermes et al., 2013) to suprasegmental properties, such as prosody (e.g., Krivokapić & Byrd, 2012) or speaking rate (e.g., Mefferd et al., 2019). Furthermore, it is a technique suitable for a range of participants, including not only healthy adult speakers but also children (e.g., Murdoch et al., 2013) and clinical populations (e.g., Mücke et al., 2018).

However, there is a large variety in approaches to EMA data collection taken by different research groups as well as some inherent differences between the different EMA devices, which may also have consequences for study findings. In my talk, I will discuss these practical aspects of EMA, underpinned by findings from our review paper (Rebernik et al., 2021a), our study comparing the accuracy of two EMA devices (Rebernik et al., 2021b), and our recent experiences with EMA data from oral cancer speakers.

Figure 1: *VOX-EMA set up in the SPRAAKLAB mobile laboratory.*

Review of EMA approaches

First, I will cover some methodological considerations of EMA as a method (including the devices in use and safety considerations), then focus on an overview of common data collection practices. This overview is based on a systematic literature review of 905 publications and includes information on common sensor placements as well as preparation methods. I will additionally describe the EMA procedure used by our research group (see Figure 1 for our current EMA setup).

VOX vs. WAVE accuracy measurement

The Northern Digital Inc. WAVE is a popular and broadly used EMA system, originally released in 2009. Its successor, NDI VOX, was released (and also discontinued) in 2020. In the second part of the talk, I will discuss our VOX vs. WAVE accuracy study (Rebernik et al., 2021b), which showed that the VOX is significantly more precise than its predecessor (see Figure 2).

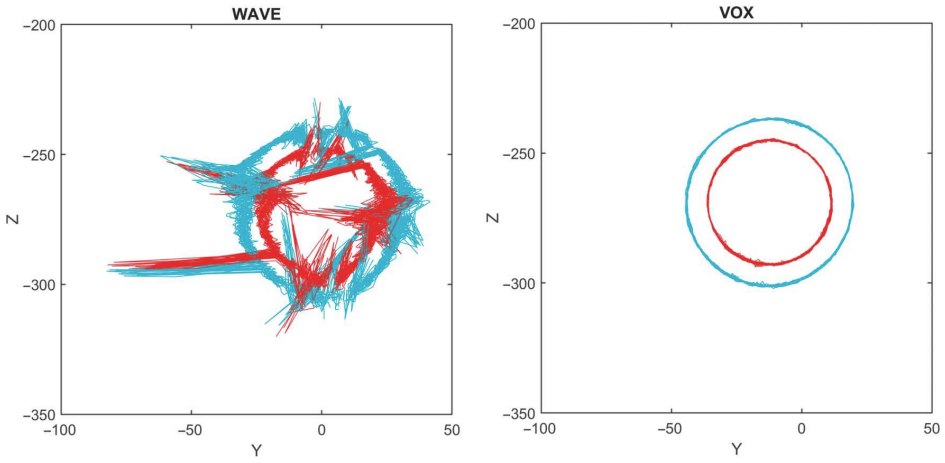


Figure 2: *Accuracy comparison of the VOX (right) vs. the WAVE (left). Figure originally published in Rebernik et al. (2021b).*

VOX-EMA data collection experience with oral cancer speakers

Finally, we will consider some practical aspects of EMA data collection. Our group has recently concluded the first study with the VOX-EMA. In a collaborative project with the Dutch Cancer Institute and the University Medical Center Groningen, we collected data from both healthy speakers as well as speakers who had undergone surgery for oral cancer (either a glossectomy or mandibulectomy). The unique anatomical characteristics of this population posed a challenge in regards to sensor placement as well as provided a learning opportunity for working with the VOX.

References

Hermes, A., Mücke, D., & Grice, M. (2013). Gestural coordination of Italian word-initial clusters: The case of “impure s.” *Phonology, 30*(1), 1–25. DOI: 10.1017/S095267571300002X

Höhne, J., Schönle, P., Conrad, B., Veldschoten, H., Wenig, P., Faghouri, H., Sandner, N., & Hong, G. (1987). Direct measurement of vocal tract shape – articulography. In *Proceedings of the European Conference on Speech Technology*, 2230–2232.

Katz, W. F., Mehta, S., & Wood, M. (2017). Using electromagnetic articulography with a tongue lateral sensor to discriminate manner of articulation. *The Journal of the Acoustical Society of America, 141*, EL57. DOI: 10.1121/1.4973907

Krivokapić, J., & Byrd, D. (2012). Prosodic boundary strength: An articulatory and perceptual study. *Journal of Phonetics, 40*(3), 430–442. DOI: 10.1016/j.wocn.2012.02.011

Mefferd, A., Efionayi, L., & Mouros, S. (2019). Tongue- and jaw-specific response patterns to speaking rate manipulations. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 3706-3710).

Mücke, D., Hermes, A., Roettger, T. B., Becker, J., Niemann, H., Dembek, T. A., … Barbe, M. T. (2018). The Effect of Thalamic Deep Brain Stimulation on Speech Production in Patients with Essential Tremor. *PLoS ONE, 13*(1), e0191359. DOI: 10.1371/journal.pone.0191359

Murdoch, B. E., Cheng, H. Y., & Barwood, C. H. S. (2013). Electromagnetic articulographic assessment of articulatory kinematics in children, adolescents, and adults. *Speech, Language and Hearing, 16*(2), 68–75. DOI: 10.1179/2050571X13Z.0000000008

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021a). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology, 12*(1), 1–42. DOI: 10.5334/labphon.237

Rebernik, T., Jacobi, J., Tiede, M., & Wieling, M. (2021b). Accuracy assessment of two electromagnetic articulographs: Northern Digital Inc. WAVE and Northern Digital Inc. VOX. *Journal of Speech, Language, and Hearing Research, 64*, 2637–2667.

Bayesian Models in the Language Sciences

Pavel Logačev, Bogazici University
pavel.logacev@gmail.com

In the last two decades, Bayesian methods have become increasingly more common in many empirical disciplines. More recently, such methods have grown in popularity in the language sciences as well. In this tutorial, I will provide a hands-on introduction to Bayesian modeling in R using the R packages brms (Bürkner, 2017), rstan (Stan Development Team, 2022) and tidybayes (Kay, 2022).

The first session will provide a brief introduction to the key ideas behind Bayesian thinking, as well as a number of worked examples in the generalized linear model (GLM) framework, including (generalized) linear mixed effects models. It will cover topics such as the specification of prior distributions, model comparison (including leave-one-out cross-validation, Bayes factors, and ROPE), as well as their interaction with variable transformations.

The second session will focus on the use of mixture models and other "custom" models outside the GLM framework which often better represent the data-generating process, as well their value in modeling behavioural and acoustic data, inter alia.

The tutorial will assume some familiarity with contemporary statistics, including linear mixed effects models.

References

Bürkner, P.C. (2017). *brms: An R package for Bayesian multilevel models using Stan*. Journal of Statistical Software, 80, 1-28.

Stan Development Team (2022). *RStan: the R interface to Stan.* https://mc-stan.org/.

Kay, M. (2022). *tidybayes: Tidy Data and Geoms for Bayesian Models.* doi:10.5281/zenodo.1308151

Recent advances of the articulatory synthesizer VocalTractLab

Peter Birkholz, Institute of Acoustics and Speech Communication, TU Dresden
peter.birkholz@tu-dresden.de

This talk provides an overview of the open-source articulatory speech synthesizer VocalTractLab (www.vocaltractlab.de) and highlights some recent advances that enhance the naturalness of the synthetic speech and allow controlling it on the level of phoneme sequences. The synthesizer consists of multiple models, namely an articulatory model of the vocal tract (Birkholz, 2013), a geometric model of the vocal folds (Birkholz et al., 2019), a one-dimensional aerodynamic-acoustic simulation (Birkholz, 2005), and a gestural control model (Birkholz, 2007). The common interface between the geometric models of the vocal tract and vocal folds on the one hand, and the acoustic simulation on the other hand is a branched tube model of the vocal system. Currently, the synthesizer comes with one “speaker model” that defines the MRI-based anatomy of a male German speaker as well as the vocal tract shapes for the context-dependent articulation of the German consonants and vowels. Future plans include the creation of models for additional speakers and other languages.
An important new feature of the recent version 2.3 of the synthesizer is the creation of the articulatory gestures (and hence the speech signal) on the basis of a given sequence of phonemes along with their durations. This mapping has been implemented based on a number of assumptions about the gestural coordination during continuous speech. Compared to the laborious manual creation of gestural scores for a synthetic utterance, the new feature allows the faster creation of synthetic utterances in consistently high quality. Additional improvements of the recent version 2.3 of the synthesizer include a reduced set of control parameters of the vocal tract model and an improved noise source model for the generation of plosives and fricatives. The talk will also discuss the synthesis of the secondary German diphthongs, which revealed the need of two (instead of one) vocal tract targets for the Tiefschwa sound (Stone et al., 2022). Finally, some applications of the articulatory synthesizer are discussed, including the simulation of early vocal learning (van Niekerk et al., 2020) and the creation and controlled manipulation of stimuli for perception experiments (Birkholz et al., 2017).

References

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. Logos Verlag, Berlin

Birkholz, P. (2007). Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In Proc. of the *Interspeech 2007 - Eurospeech*, pp. 2865–2868, Antwerp, Belgium

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4): e60603

Birkholz, P., Drechsel, S., Stone, S. (2019). Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis. In *Proc. of the Interspeech 2019*, pp. 3765-3769, Graz, Austria

Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., Neuschaefer-Rube, C. (2017). Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis. *Computer Speech & Language*, 41, pp. 116-127

Stone, S., Gao, Y., Birkholz, P. (2022). Articulatory synthesis of vocalized /r/ allophones in German. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 879-889

van Niekerk, D.R., Xu, A., Gerazov, B., Krug, P.K., Birkholz, P., Xu, Y. (2020). Finding intelligible consonant-vowel sounds using high-quality articulatory synthesis. In *Proc. of the Interspeech 2020*, pp. 4457-4461, Shanghai, China

Task Dynamic Application

Dzhuma Abakarova, ZAS, University of Potsdam
abakarova@leibniz-zas.de

Speech production is a complex phenomenon that involves a variety of cognitive and physiological processes many of which are inaccessible to direct investigation. Computational models can play an important role in advancing our understanding of the mechanisms involved in speech. Modelling provides a rigorous way of summarizing the available knowledge, generates new predictions, and forces us to be very specific in our theoretical assumptions. Importantly, it allows for testing hypothesized causes for observed regularities in speech production kinematics when it is either impossible or impractical to create experimental conditions for direct measurement.

Task Dynamic Application (TaDA, Nam et al., 2004) is an example of such model. TADA is a computer implementation of Articulatory Phonology framework (Browman & Goldstein, 1992) that combines the Task Dynamic model (Saltzman & Munhall, 1989), a coupled-oscillator model of inter-gestural planning (Goldstein et al., 2009), and a gestural-coupling model. It models both planning and execution stages of speech production, from gestures and their overlap to articulatory motions and acoustics. The model's editing capabilities allow for implementation and testing of hypotheses about the effects of amount of gestural overlap, gestural parameters, prosodic or speaker differences on articulation, and many more. The simulations result into articulatory and acoustic trajectories that can be compared to experimental data. As an articulatory-acoustic model, TaDA can also be used to generate perceptual stimuli with known (pseudo-)articulatory properties. TADA allows for the development of articulatory models for different languages through the use of language-specific dictionary files and gestural databases.

In this tutorial, I will introduce the Task Dynamic Application, the theoretical framework behind it, its components and architecture, the inputs and outputs. I will also give an overview of its potential uses with reference to previous studies that have used TaDA. Finally, we are going to conduct a small simulation experiment that will provide a hands-on experience with using TADA. The goal of the tutorial is to give its participants an idea of the possibilities (and limitations) of TADA as a research tool as well as the requirements and skills necessary for using it.

References

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled Oscillator Planning Model of Speech Timing and Syllable Structure. In C. G. M. Fant, H. Fujisaki, & J. Shen (Eds.), *Frontiers in phonetics and speech science* (p. pp.239-249). The Commercial Press. https://hal.archives-ouvertes.fr/hal-03127293

Nam, H., Goldstein, L., Saltzman, E., & Byrd, D. (2004). TADA: An enhanced, portable Task Dynamics model in MATLAB. *The Journal of the Acoustical Society of America*, *115*(5), 2430. https://doi.org/10.1121/1.4781490

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Public engagement in speech research: some illustrations and potential benefits

Martijn Wieling, University of Groningen, Speech Lab Groningen
m.b.wieling@rug.nl

In this session, I will illustrate approaches which we – the Speech Lab Groningen (SLG) team – have used to engage with a general audience and involve them in our research. While I will illustrate initiatives which we have used, the session is also explicitly meant for sharing experiences, so that participants may inspire each other.

The research in the SLG focuses on several different strands of (speech) research:

1) Quantitatively investigating variation and change with a focus on regional languages (e.g., Wieling and Nerbonne, 2015; Buurke et al., 2021);
2) Developing language and speech technology for regional languages (e.g., Bartelds & Wieling, 2022; Bartelds et al., 2022; de Vries et al., 2021);
3) Investigating the speech articulation of several populations, often those with disordered speech due to (e.g.,) Parkinson's disease, oral cancer (e.g., Wieling, 2018; Jacobi et al., 2020; Rebernik et al., 2021).

Our public engagement activities often showcase the research in several of these research lines. One aspect of these activities is to create output aimed at a general audience. Some examples include a comic about an electromagnetic articulography study (Wieling et al., 2016; see Figure 1), a board game (see Figure 2) in which players need to spread the regional language they represent to other regions (similar to the board game Risk, but instead of rolling the dice, players need to answer dialect/linguistics questions), and also a digital app (iOS and Android: "Van Old noar Jong"), with an associated educational program, to teach primary school children aspects of the local regional language (i.e. *Gronings*).

Figure 1: Comic (full version: www.martijnwieling.nl/comic)

Figure 2: Board game "Streektaalstrijd"

Figure 3: SPRAAKLAB

While this output is one-way, we also try to engage with the general audience directly. We do this, for example, by interacting with them during science festivals and other events. To interest people in our research, we demonstrate various techniques. For example, we use an ultrasound tongue imaging device to show children (and adults) the movement of their tongue during speech. As a keepsake, we also give children a printout from the ultrasound image of their tongue. Furthermore, we combine an electrolanyx with the VTM-S20 vocal tract model of Prof. Takayuki Arai (Arai, 2020) and invite children and adults generate different vowels. Of course, we also bring our comic, demonstrate our *Gronings* app using tablets, etc.

To make public engagement also directly beneficial for our own research, we often collect scientific data at public engagement events. For example, we have developed an app which predicts where someone is from on the basis of their recorded (dialectal) speech (cf. Hilton, 2021). In addition, for two consecutive years, we were selected for *Lowlands Science*, through which visitors of the large 60,000-visitor three-day music festival *Lowlands* can participate in (attractive) scientific research. In 2018, we have investigated the influence of alcohol on native and non-native speech on the basis of data from about 250 participants (Offrede et al., 2021), while in 2019 we have investigated phonetic convergence in speech using a novel paradigm (Wieling et al., 2020). In addition, whenever we are able to bring our mobile laboratory SPRAAKLAB (Figure 3) to an event, we often collect data using short experiments (in our sound-dampened room). For example, during the summer of 2021, we have collected data using a formant perturbation experiment during six days of the *Noorderzon* festival.

In sum, in our lab, we do not only view public engagement activities as a very enjoyable aspect of our work as researchers, but we also view it as an opportunity to directly benefit our own research by collecting data. An added benefit of our public profile for me as a lab coordinator is that students who volunteer to help out during our public engagement activities, frequently remain active for our lab, and have sometimes even ended up as PhD students in the SLG. Consequently, strengthening our lab's public engagement profile has certainly strengthened our lab's scientific profile.

References

Arai, T. (2020). Acoustic-phonetics demonstrations for classroom teaching. *The Journal of the Acoustical Society of America*, 148(4), 2609-2609.

Bartelds, M., & Wieling, M. (2022). Quantifying language variation acoustically with few resources. *Proceedings of NAACL*.

Bartelds, M., de Vries, W., Sanal, F., Richter, C., Liberman, M., & Wieling, M. (2022). Neural representations for modeling variation in speech. *Journal of Phonetics*.

Buurke, R., Sekeres, H., Heeringa, W., Knooihuizen, R., & Wieling, M. (2021). Estimating the level and direction of phonetic dialect change in the northern Netherlands. *arXiv preprint* arXiv:2110.07918.

de Vries*, W., Bartelds*, M., Nissim, M., & Wieling, M. (2021). Adapting monolingual models: data can be scarce when language similarity is high. *Findings of ACL*, pp. 4901-4907. arXiv:2105.02855v2

Hilton, N. H. (2021). Stimmen: A citizen science approach to minority language sociolinguistics. *Linguistics Vanguard*, 7(s1).

Jacobi, J., Rebernik, T., Jonkers, R., Maassen, B., Proctor, M., & Wieling, M. (2020). Characterizing tongue tremor in Parkinson's disease using EMA. *Proceedings of the 12th International Seminar on Speech Production*, pp. 80-83.

Offrede, T., Jacobi, J., Rebernik, T., de Jong, L., Keulen, S., Veenstra, P., Noiray, A., & Wieling, M. (2021). The Impact of Alcohol on L1 vs. L2. *Language and Speech*, 64(3), 681-692. doi: 10.1177/0023830920953169.

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, 12(1), 6. doi: 10.5334/labphon.237.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetic*s, 70, 86-116.

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. Annual Review of Linguistics, 1, 243-264.

Wieling, M., Tiede, M., Rebernik, T., de Jong, L., Braggaar, A., Bartelds, M., Medvedeva, M., Heisterkamp, P., Freire Offrede, T., Sekeres, H., Pot, A., van der Ploeg, M., Volkers, K., & Mills, G. (2020). A novel paradigm to investigate phonetic convergence in interaction. *Proceedings of the 12th International Seminar on Speech Production*, pp. 1-4.

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, 59, 122-143.

Regional Dialect Influence on L2 Voiced Affricate Production: The case of L1 Turkish and L2 English

Bahar Aksu, Lancaster University
b.aksu@lancaster.ac.uk

Second Language speech models, such as L2LP (Escudero, 2005), and SLM-r (Flege & Bohn, 2021) suggest that differences between speakers in their L1 can influence patterns of L2 speech. That is, fine grained phonetic differences that exists between L1 dialects may explain variation in those speakers' L2 speech. While studies of L2 speech perception have demonstrated an influence of L1 regional dialect on L2 speech perception (Chládková & Podlipský, 2011, Escudero & Williams, 2012), the few studies on the effects of L1 dialect on L2 speech production have shown more mixed findings (Marinescu, 2012, Simon et al., 2015). This study aims to advance our understanding of L2 speech production by investigating L2 English vowel production among L1 speakers of Turkish who speak one of the two Turkish regional dialects, namely İstanbul and Trabzon. Specifically, I investigate the L2 English voiced affricate production of İstanbul and Trabzon Turkish speakers. Similar to English, the phonetic realization of the voiced affricate is palato-alveolar in İstanbul Turkish. However, Brendemoen (2002) states that dental alveolar is a salient feature of the eastern part of the Trabzon Turkish. Thus, this study aims to examine whether tis regional variation would lead dialect speakers to differ in their realization of the voiced affricate in L2 English.

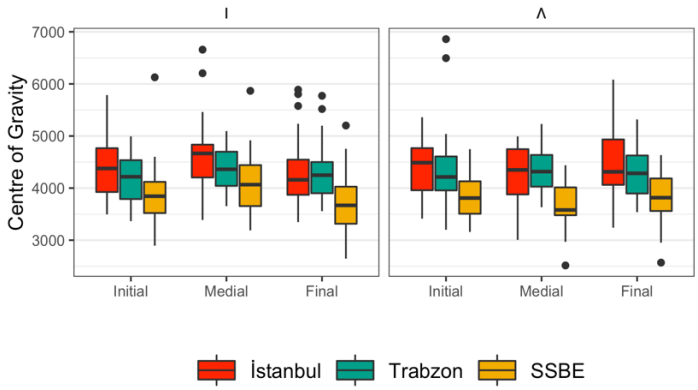


Figure 1: Centre of Gravity among dialect speakers in word position and the vowel context

Data was collected in Turkey and in the UK from Turkish-English bilinguals and Standard Southern British English (SSBE) speakers. A language background questionnaire and a wordlist in English were presented to 14 L1 Turkish speakers from each dialect and 14 SSBE speakers aged 18 -35 (N=42). Two tokens for voiced affricate at word-initial, word-medial, and word-final position were examined for acoustic analysis. Centre of Gravity, Skewness, Kurtosis, duration of the preceding vowel, and the duration of the frication were measured in Praat. Mixed-Effects models were fit to observe regional dialect influence on L2 voiced affricate production. The results revealed that there was not a significant effect of regional dialects on the acoustic correlates of L2 voiced affricate produced by Trabzon and İstanbul speakers. However, regional dialect is found to be interacted with the word position and the preceding vowel. While the similarity is greater at word-initial and word-final position among speakers, regional dialect was found to be interacted with word-medial position. In addition, the results have shown that the acoustic realization of the voiced affricate is significantly different between L1 Turkish and SSBE speakers, even though the phonetic categorization of this consonant is the same for both languages (See Figure 1). These findings will be discussed according to framework of SLM-r in L2 speech production.

References

Brendemoen, B. (2002). The Turkish Dialects of Trabzon. Their Phonology and Histrocial Development I-II. [Turkologica, Band 50] Wiesbaden, Harrassowitz Verlag.

Boersma, P. (2019). Praat: doing phonetics by computer. [Computer Prgram]. Version 6.1

Chládková, K., & Podlipský, J.V. (2011). Native dialect matters: Perceptual assimilation of Dutch vowels by Czech listeners. *The Journal of Acoustical Society of America,* 130(4), EL186 -EL 192. https://doi.org/10.1121/1.3629135

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. (PhD thesis, University of Utrecht)

Escudero, P., Williams, D. (2012). Native dialect influences second language vowel perception: Peruvian versus Iberian Spanish learners of Dutch. *The Journal of Acoustical Society of America*, 131(5), EL406 - EL412.

Flege, J.E., & Bohn, O.S. (2021) The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3-83). Cambridge University Press

Marinescu, I. (2012) Native dialect effects in non-native production and perception of vowels (PhD Thesis, University of Toronto)

Simon, E., Debaene, M., & Van Herreweghe, M. (2015). The effects of L1 regional variation on the perception and production of L1 and L2 vowels. *Folia Linguistica*, 49(2), 521-553. https://doi.org/10.1515/flin-2015-0018

Effect of orofacial somatosensory stimulation in speech perception in relation to speech production performance

Monica Ashokumar[1], Clément Guichet[1], Jean-Luc Schwartz[1], Takayuki Ito[1,2]
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France, [2] Haskins Laboratories, New Haven, USA
monica.ashokumar@gipsa-lab.grenoble-inp.fr, clement.guichet@etu.univ-grenoble-alpes.fr, jean-luc.schwartz@gipsa-lab.grenoble-inp.fr, takyuki.ito@gipsa-lab.grenoble-inp.fr

Orofacial somatosensory inputs modify the perception of speech sounds (Ito et al., 2009, Trudeau-Fisette et al., 2019). This indicates that the somatosensory information associated with speech production can be involved in the processing of speech perception through sensory-motor links. Accordingly, this somatosensory-auditory interaction mechanism can be developed along with the acquisition of speech production. However, it is still unknown whether the effect of somatosensory inputs in speech perception is related to characteristics or level of speech production ability. We here investigated whether the somatosensory effect in vowel perception can be varied depending on the individual characteristics of corresponding vowel production. Somatosensory effect in speech perception was examined using a vowel identification test between French vowels /e/ (unrounded) and /ø/ (rounded) in an 8-member /e/-/ø/ continuum as in Trudeau-Fisette et al., (2019). When presenting a stimulus sound in the vowel identification test, orofacial somatosensory stimulation associated with facial skin deformation was also applied using a robotic device in backwards direction, which can correspond to the articulatory movement for the production of the vowel /e/. The perceptual boundary between /e/ and /ø/ was determined by estimating psychometric function. The somatosensory effect was quantified as a difference in perceptual boundary between two conditions with and without somatosensory stimulations. Speech production characteristics were assessed based on the first, second and third formants (F1, F2, and F3) of the vowels /e/ and /ø/. Those vowels were recorded by the same participants in a separate recording session. Differences between /e/ and /ø/ in each formant were used as production index. We confirmed that orofacial somatosensory stimulation altered vowel categorization and significantly increased the amount of /e/ responses as found in Trudeau-Fisette et al. (2019). We found that amplitude of somatosensory effect in perception was correlated with speech production index in F2 and F3, but not in F1. Since acoustical contrast between /e/ and /ø/ is characterized mostly in F2 and possibly in F3, but rarely in F1, the results showed a clear relationship between somatosensory effect and production ability. This could support the idea that the development of somatosensory function in speech perception can be related to the acquisition of speech production.

References

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. Proceedings of the National Academy of Sciences, 106(4), 1245–1248. https://doi.org/10.1073/pnas.0810063106

Trudeau-Fisette, P., Ito, T., & Ménard, L. (2019). Auditory and Somatosensory Interaction in Speech Perception in Children and Adults. Frontiers in Human Neuroscience, 13, 344. https://doi.org/10.3389/fnhum.2019.00344

Exploring initial-consonant lengthening in typologically diverse languages with Bayesian linear regression

Frederic Blum, HU Berlin & Max-Planck Institute for Evolutionary Anthropology
frederic_blum@eva.mpg.de

This study explores the lengthening of initial consonants across 25 typologically diverse languages using Bayesian linear regression. Special attention is given to the variation of the effect, both between languages as well as between speakers of a certain language. The data used in this study comes from the DoReCo project (Seifart et al., 2022) and includes semi-automatically time-aligned phonemes (Paschen et al., 2020).

Acoustic lengthening of segments before and after domain-boundaries supports the segmentation of different elements of speech (Klatt, 1976; Beach, 1991; Fletcher, 2010; Price et al., 1991; Shatzman and McQueen, 2006). While earlier explanations focused on speech planning as an explanation for acoustic lengthening effects (Cooper and Paccia-Cooper, 1980), most if not all scholars nowadays agree that segmental lengthening effects do not arise from physiological or cognitive constraints of the speaker alone (Fletcher, 2010; Katz and Fricke, 2018; White, 2014). There is a lot of evidence suggesting that those effects serve a communicative function for the listener, namely facilitating the processing of different linguistic structures (Oller, 1973; Turk and Shattuck-Hufnagel, 2000). For example, lengthening has been shown to provide cues for the segmentation of lexical elements for speakers of various languages, such as Dutch and French (Christophe et al., 2004), English (Cho et al., 2007), Korean (Kim et al., 2012), Swedish (Lindblom, 1968), and Warlpiri (Butcher and Harrington, 2003). Despite this low coverage of languages and the wide the range of diversity across languages and language families (Evans and Levinson, 2009) as well as between speakers (Yu and Zellou, 2019), some studies even go so far as to claim universality for the effect of word-initial consonant lengthening (White et al., 2020).

I want to contribute to this discussion by exploring the effect of acoustic lengthening in a wide range of languages using Bayesian methods. From this decision follows that all results will include explicit mentions of uncertainty (McElreath, 2020; Vasishth and Gelman, 2021). The goal is not to confirm or reject a certain hypothesis, but to enrich the discussion on lengthening at domain-edges by a typological perspective, and to explore the effect of initial-consonant lengthening in structurally and phylogenetically diverse languages. In order to tackle the problem of generalizability beyond the sample (Winter and Grice, 2021; Yarkoni, 2020), a focal point of this study is the analysis of varying effects. The target effect, divided into the levels utterance-initial, word-initial and non-initial, is analyzed with varying slopes and intercepts across different predictors, which allows for studying the variation within multiple categories. Especially the variation between languages (Evans and Levinson, 2009) and speakers (Yu and Zellou, 2019) has been shown to be relevant. Only by considering this possibility of varying effects across all those categories will we be able to start making generalizing claims about the lengthening of initial consonants.

References

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30 (6), 644–663. https://doi.org/10.1016/0749-596X(91)90030-N

Butcher, A., & Harrington, J. (2003). An acoustic and articulatory analysis of focus and the word/morpheme boundary distinction in Warlpiri. *Proceedings of the 6th International Seminar on Speech Production*, 19–24.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, 35 (2), 210–243. https://doi.org/10.1016/j.wocn.2006.03.003

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, 51 (4), 523–547. https://doi.org/10.1016/j.jml.2004.07.001

Cooper, W. E., & Paccia-Cooper, J. (1980). *Syntax and Speech*. Harvard University Press. https://doi.org/10.4159/harvard.9780674283947

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32 (5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 521–602). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444317251.ch15

Katz, J., & Fricke, M. (2018). Auditory disruption improves word segmentation: A functional basis for lenition phenomena. *Glossa: a journal of general linguistics*, 3 (1), 38. https://doi.org/10.5334/gjgl.443

Kim, S., Cho, T., & McQueen, J. M. (2012). Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *Journal of Phonetics*, 40 (3), 443–452. https://doi.org/10.1016/j.wocn.2012.02.005

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59 (5), 1208–1221. https://doi.org/10.1121/1.380986

Lindblom, B. (1968). Temporal organization of syllable production. *Quarterly progress and status report*, 9 (2-3), 1–5.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC press.

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of the Acoustical Society of America*, 54 (5), 1235–1247. https://doi.org/10.1121/1.1914393

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., & Seifart, F. (2020). Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo). *Proceedings of The 12th Language Resources and Evaluation Conference*, 2657–2666. https://www.aclweb.org/anthology/2020.lrec-1.324

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90 (6), 2956–2970. https://doi.org/10.1121/1.401770

Seifart, F., Paschen, L., & Stave, M. (2022). *Language Documentation Reference Corpus (DoReCo)*.

Shatzman, K. B., & McQueen, J. M. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception & Psychophysics*, 68 (1), 1–16.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. J*ournal of Phonetics*, 28 (4), 397–440.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, 59 (5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

White, L. (2014). Communicative function and prosodic form in speech timing. *Speech Communication*, 63, 38–54. https://doi.org/10.1016/j.specom.2014.04.003

White, L., Benavides-Varela, S., & Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics*, 81, 100982. https://doi.org/10.1016/j.wocn.2020.100982

Winter, B., & Grice, M. (2021). Independence and generalizability in linguistics. *Linguistics*, 59 (5), 1251–1277. https://doi.org/10.1515/ling-2019-0049

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, 5, 131–150. https://doi.org/10.1146/annurev-linguistics-011516-033815

Reduced forms in conversation: item and speaker characteristics

Kübra BODUR, Aix-Marseille Université
kubra.bodur@univ-amu.fr

In spoken language, a significant proportion of words are produced with missing or underspecified phonetic forms which is a process called phonetic reduction (Johnson, 2004 ; Ernestus & Warner, 2011). Phonetic segments may be weakened or completely absent in some cases. Our study aims to better understand how this complex production process works and to highlight the relationship between phonetic reduction and two possible factors of variation: a/ the nature of the items themselves, considering that all items are not systematically affected in the same way by reduction; b/ the specificity of the speakers, which is not often mentioned as a possible factor in the appearance of reduced forms.

With that purpose, we analyzed 13 reduced forms (words and word combinations) frequently encountered in everyday French (e.g., Adda-Decker & Snoeren, 2011 ; Wu & Adda-Decker, 2020). Using an Enriched Orthographic Transcription for the extraction of reduced items, we observed their distribution and duration in a corpus of spontaneous/conversational speech (Bertrand et al., 2008).

The results show that the frequency of a reduced item is quite heterogenous across conversations and that it depended on the item itself as well as on the speaker. Although all speakers produce reduced forms, some speakers reduce more often than the others I the corpus. Post-hoc analyses revealed a weak correlation ($p= 0.4203$) between the rate of reduction of each speaker and his/her speech rate (phonemes per second) in the corpus. Finally, an analysis on 3 items (*alors, enfin, parce que*) shows that, once again, each item was not reduced consistently in the same forms by all speakers, suggesting that each speaker has his/her own reduction patterns.

References

Adda-Decker, M., & Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 39(3), 261-270.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Revue TAL*, 49(3), 105-134.

Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(SI), 253-260.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis*. Proceedings of the 1st session of the 10th international symposium (pp. 29-54).

Wu, Y., & Adda-Decker, M. (2020). Réduction temporelle en français spontané: où se cache-t-elle? Une étude des segments, des mots et séquences de mots fréquemment réduits. In *Actes des 33èmes Journées d'Etudes sur la Parole*, June 2020, Nancy, France (pp. 627-635).

Do surprisal and background noise affect speech clarity in the same way? From production to perception

Omnia Ibrahim, Department of Language Science and Technology, Saarland University, Germany
omnia@lst.uni-saarland.de

Speakers modulate the characteristics of their own speech and produce a listener- oriented, clear speaking style in response to communication demands (Bradlow, 2002; Lindblom, 1990). Such clear speech often takes the form of increased loudness, higher pitch, expanded vowel space, hyper-articulation, and lengthening, with perceptual consequences of improved intelligibility. Other causes could also lead to clear speech: context predictability (Aylett and Turk, 2006) or presence of background noise, i.e. the Lombard effect (Lu and Cooke, 2009). Since both context predictability and the Lombard effect contribute to enhance acoustic signals, it is not clear whether or not these two effects contribute to the signals in the same way (Q1) and if such contributions will affect speech clarity in perception (Q2). Results from a production experiment will be reported in the current study, followed by that of a perception experiment, which is still in progress.

A total of 1520 target CV syllables were annotated and analysed from 38 German speakers. The stimuli were recorded in two conditions no noise and -10 dB white-noise SNR with 2 surprisal contexts (High vs. Low). Surprisal, defined as $S(\text{syllable}_i) = -\log_2 P(\text{syllable}_i|\text{Context})$, was estimated by means of a syllable- level language model trained on DeWaC (Kilgarriff et.al., 2010). We measured acoustic features extracted from our target syllables (duration, intensity (average and range) and median fundamental frequency) (See figure1) and from the vowel (F1, F2 and F2-F1 distance) (See figure2) inside the syllable. The analysis's results revealed a significant effect of presence vs. absence of noise on all the measured features (except for the second formant where we found the effect of noise on front vowels only), while we found an effect of surprisal on syllable duration and intensity range only. Overall the effects were additive, not interactive. These findings suggest that noise-related modifications are independent from predictability-related changes, with implications for including channel-based and message-based formulations in speech production models. In the perception experiment, to investigate the contribution of those acoustic enhancements in speech perception, listeners were instructed to discriminate a range of target syllables with varying degree of clarity.

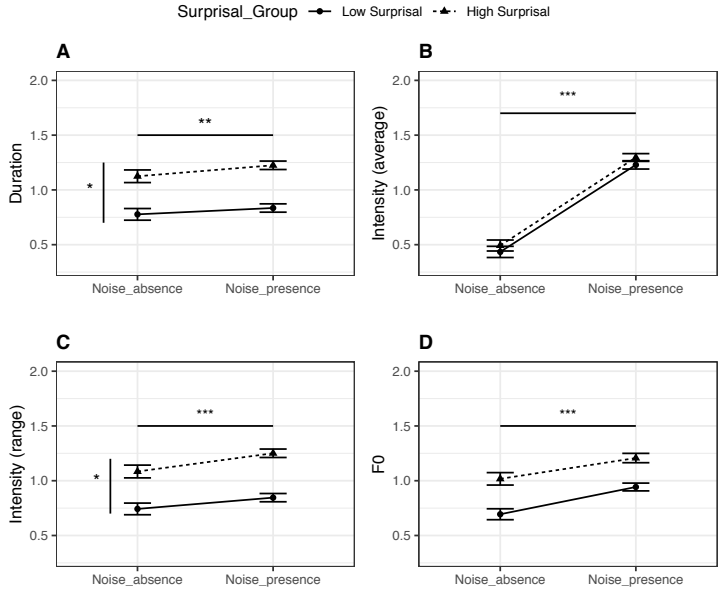


Figure 1: Mean z-scores for syllable-level (A) duration, (B) average intensity, (C) intensity range and (D) F0 as a function of absence vs. presence of noise.

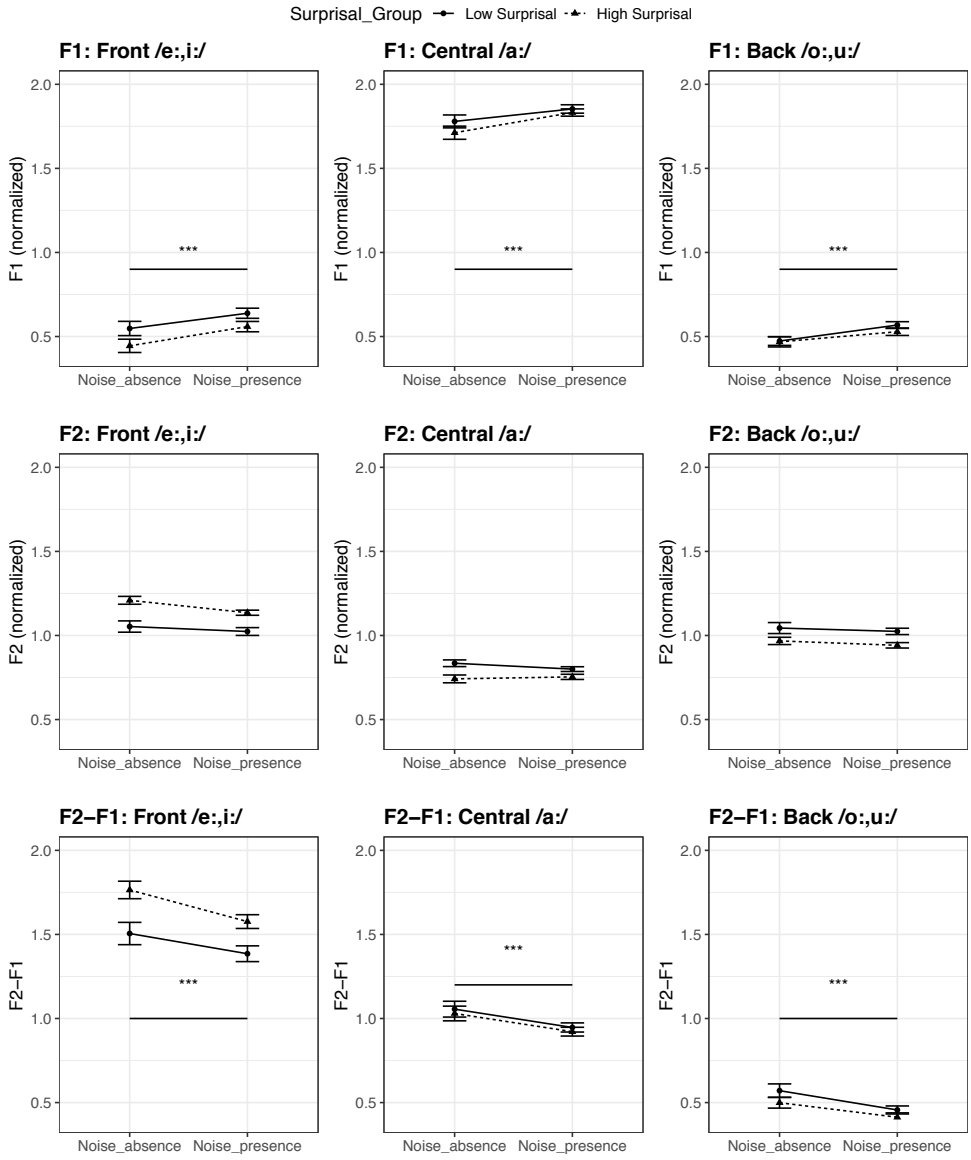


Figure 2: Mean z-scores of F1, F2 measured at the mid-point of the vowel, and F2-F1 difference for front /i:, e:/, central /a:/ and back /o:, u:/ vowels in the absence vs. presence of noise.

References

Aylett, M.; Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America,* 119, 3048–3058.

Bradlow, A. R. (2002). Confluent talker- and listener-related forces in clear speech production. *Laboratory phonology* 7, C. Gussenhoven and N. Warner (Ed.), (pp. 241–73). Berlin, Germany/New York, NY: Mouton de Gruyter.

Kilgarriff, A.; Reddy,S. ; Pomikálek, J.; Avinesh, P.V.S. (2010). *A corpus factory for many languages*. LREC workshop on Web Services and Processing Pipelines, Malta, May 2010.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H & H theory (pp. 403–439). Kluwer Academic Publishers.

Lu, Y.; Cooke, M (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Communication*, 51, 1253-1262.

Pharyngealization across different contexts: A Bayesian account

Philipp Buech, Laboratoire de Phonétique et Phonologie – UMR 7018 (CNRS/Sorbonne Nouvelle)
philipp.buech@sorbonne-nouvelle.fr

Pharyngealization is a secondary articulation which is produced by a backing of the tongue (Trask, 1996) and it is a typologically rare but prominent feature of some Afro-Asiatic languages. The acoustic characteristics are well-described for Arabic varieties (see, e.g., Al-Tamimi, 2017; Kulikov, 2022; Shar and Ingram, 2010), showing that pharyngealization is primarily cued by a lower F2 of the following vowel. Some articulatory studies exist also, having a description of a backed tongue root in common (see, e.g., Hermes, 2018). The role of the tongue dorsum is a matter of debate, which was described as being lowered, raised or retracted (Alwabari, 2022; Embarki et al., 2011; Zeroual et al., 2011). However, most studies focus on data of CV or VCV sequences, as secondary articulations are signalled mostly by surrounding vowels (Ladefoged & Maddieson, 1996). An open question is how the production of pharyngealization is realized in (at least partial) consonantal environments such as [C_V], [V_C], and [C_C].

This study aims to provide an analysis of the acoustic and articulatory patterns of the plain/ pharyngealization distinction in Tashlhiyt Amazigh: [V_V] (e.g., [ada/adˤa]), [C_V] (e.g., [bda/bdˤa]), [V_C] (e.g., [adg,adˤg]), and [C_C] (e.g., [bdg,bdˤg]). Tashlhiyt Amazigh is a language spoken in North Africa (Chaker & Mettouchi, 2009), where pharyngealization is a contrastive feature in the set of coronals and which is characterized by strong use of consonant clusters. The analysis of the data was carried out entirely in the Bayesian framework.

Articulatory and acoustic data from six native male speakers of Tashlhiyt were collected using the electromagnetic articulograph AG 501 (Carstens Medizinelektronik GmbH, 2014). Sensor coils were placed on the upper and lower lips (ULIP, LLIP), tongue tip (TTIP), tongue mid (TMID), and the tongue body (TBO), with additional reference sensors behind the left and the right ear. The speech material consisted of words containing the plain/pharyngealized target consonants [d] and [dˤ] in the environments [a_a], [b_a], [a_g], [b_g], that were entered into a carrier sentence, e.g., *Innayam **adan** bahra*, 'He told you (fem) intestine a lot'. Acoustic and articulatory measurements were automatically retrieved with a script written in Python v.3.10.4 (Rossum & Drake, 2009) and using Parselmouth v. 0.4.0 (Jadoul et al., 2018). PyMC3 v3.11.2 (Salvatier et al., 2016) was used for the Bayesian statistical analysis.

In the articulation, the trajectories of the LLIP, TTIP, TMID and TBO sensors (low-high and frontback dimensions) were extracted from the acoustic start to the acoustic end of the target sequences. In the acoustics, the trajectories of the formants F1, F2, and F3 of the surrounding vowels were extracted. Both, the kinematic and formant trajectories of the plain and the pharyngealized condition will be compared over time and across contexts. Additional stationary consonant-related acoustical measurements will also be presented.

References

Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8(1), 1–40. https://doi.org/10.5334/labphon.19

Alwabari, S. (2022). *Phonological and Physiological Constraints on Assimilatory Pharyngealization in Arabic: Ultrasound Study* (PhD dissertation). University of Ottawa. Retrieved March 3, 2022, from https://ruor.uottawa.ca/handle/10393/40302

Carstens Medizinelektronik GmbH. (2014). *AG501 manual*. Retrieved April 2, 2022, from http://www.ag500.de/manual/ag501/ag501-manual.pdf

Chaker, S., & Mettouchi, A. (2009). Berber. In K. Brown & S. Ogilvie (Eds.), *Concise Encyclopedia of the Languages of the World* (pp. 152–158). Elsevier.

Embarki, M., Ouni, S., Yeou, M., Guilleminot, C., & Maqtari, S. A. (2011). Acoustic and electromagnetic articulographic study of pharyngealisation. Coarticulatory effects as an

index of stylistic and regional variation in Arabic. In Z. M. Hassan & B. Heselwood (Eds.), *Instrumental studies in Arabic Phonetics* (pp. 193–216). John Benjamins Publishing Company.

Hermes, Z. (2018). *The phonetic correlates of pharyngealization and pharyngealization spread patterns in cairene arabic an acoustic and real-time magnetic resonance imaging study* (PhD dissertation). University of Illinois at Urbana-Champaign. Retrieved May 21, 2022, from http://hdl.handle.net/2142/102935

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Kulikov, V. (2022). Voice and Emphasis in Arabic Coronal Stops: Evidence for Phonological Compensation. *Language and Speech*, 65(7), 73–104. https://doi.org/10.1177/0023830920986821

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Blackwell Publishers. Rossum, G. V., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 1–24. https://doi.org/10.7717/peerj-cs.55

Shar, S., & Ingram, J. (2010). Pharyngealization in assiri arabic: An acoustic analysis. *Proceedings of the 13th Australian International Conference on Speech Science and Technology*, 5–8.

Trask, R. L. (1996). *A Dictionary of Phonetics and Phonology*. Routledge.

Zeroual, C., Esling, J. H., & Hoole, P. (2011). EMA, endoscopic, ultrasound and acoustic study of two secondary articulations in Moroccan Arabic. In *Instrumental studies in Arabic Phonetics* (pp. 277–298). John Benjamins Publishing Company.

Articulatory and Sociocultural Constraints in the Singing Voice: the pronunciation of intervocalic /t/

Caillol Coline, CLILLAC-ARP (UR 3967) Université Paris Cité
coline.caillol@u-paris.fr

In the 1980s, sociolinguist Peter Trudgill observed that some British pop artists such as The Beatles tended to adopt certain American pronunciation features in their singing voice, though they did not display them in their speaking voice (Trudgill, 1983). My work seeks to study whether this pronunciation-shift towards Americanization can also be applied to a genre such as traditional British Heavy Metal, whose inception was highly situated within the socio-cultural context of Northern England in the 1970s and 1980s, stemming from a disillusioned young generation faced with deindustrialization (Walser, 1993; Weinstein, 2000).

An extensive corpus study was conducted studying all songs (675) from all studio albums of four archetypal British Heavy Metal bands (Black Sabbath, Judas Priest, Iron Maiden and Def Leppard), as well as a certain number of selected interviews for speaking voice comparison. Multiple pronunciation features distinguishing British English from American English were analyzed, among which word-internal intervocalic /t/, which is the current focus of my dissertation and has been described as “one of the most striking characteristics of American pronunciation to the ears of a non-American” (Wells, 1982). The typical American realization of /t/ in a V_v context is voiced and commonly called a flap, whereas British English tends to realize intervocalic /t/ as a voiceless stop. Results showed that while flapping was not found in the spoken productions of the four bands, it did occur to various degrees according to band in their sung productions.

Socio-cultural explanations for this pronunciation shift can be put forward, including but not limited to, the domination of the American music industry at that time and the belief that for these bands to succeed, they had to “make it big” in the US. In a more abstract way, this work questions the relationship between identity (both projected and perceived) and music as a cultural production and performance, and how these notions tie back to that of accents. Singing is both a performance directed at an audience (Bell, 1984, 2002; Frith, 1998) as well as an expression of a particular identity. The discrepancy between the way performers present themselves (with costumes, set designs, album covers and lyrics all relating to their British origins) and their Americanized singing-voice pronunciation is what Trudgill refers to as “acts of conflicting identity.” This idea refers back to the multimodal nature of speech (and singing) and adds further challenge the already complex notion of speaker-listener interaction.

Adding to these socio-cultural considerations, it turns out that articulatory constraints specific to the act of singing may also play a role in the pronunciation shift of British Heavy Metal artists. Taking into account the central notions of ease of articulation as described in Lindblom’s Hyper and Hypo-articulatio theory (1990), singers resort to flapping most word-internal intervocalic /t/ in their songs because it may simply be easier to do so than to produce its plosive counterpart. Factors such as word frequency, tempo, pitch and the necessity to carry a smooth musical line without interruptions or changes in intensity may all potentially contribute to this argument. More specifically, I am currently planning an experiment using electropalatography (EPG) technology to evaluate how higher pitched passages (frequent in traditional British Heavy Metal) may lead to a flapped /t/ realization simply because the articulatory setting required to produce a prototypical plosive is much harder to reach with the bigger buccal opening necessary to sing at high pitches (Cornut, 2004; Scotto Di Carlo, 2005). While I do not yet have results for this experiment, it nevertheless raises some questions directly related to the theme of this winter school. EPG studies are particularly expensive, as each participant must have its own custom-made palate, and time-consuming, which poses issues of generalizability and

reproducibility of results. Furthermore, the highly situated socio-cultural and historical aspects of the singing productions under study mean that they cannot so easily be taken out of context and placed in a lab setting with the hopes of achieving similar results.

The ultimate goal of my dissertation is to establish a trade-off model of the singing voice pronunciation of British artists, quantifying factors of influence both socio-cultural (British origins, American model) and articulatory (ease of articulation, intelligibility) and looking at how they interact with one another. The multidisciplinary nature of my research, at a crossroads between sociology, cultural studies, corpus and experimental phonetics compels me to directly consider and address the complexity of speech production and perception, make the most of recent developments in the field and attempt to come up with original and innovative resources to cope with these challenges.

Bibliography

Bell, A. (1984). Language Style as Audience Design. *Language in Society*, *13*(2), 145-204.

Bell, A. (2002). Back in style: reworking audience design. In Eckert, P. and Rickford, J. R., editors, *Style and Sociolinguistic Variation*, pages 139–169. Cambridge University Press, Cambridge.

Cornut, G. (2004). *La voix*. Presses Universitaires de France.

Frith, S. (1998). *Performing Rites: Evaluating Popular Music*. Oxford University Press, Oxford, revised edition.

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Éds.), *Speech Production and Speech Modelling* (p. 403-439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Scotto Di Carlo, N. (2005). *Contraintes de production et intelligibilité de la voix chantée*. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA), 24:159–176.

Walser, R. (1993). *Running with the devil: Power, gender, and madness in heavy metal music*. Wesleyan University Press.

Weinstein, D. (2000). *Heavy metal: The music and its culture*. Da Capo Press.

Wells, J. C. (1982). *Accents of English. 1: An introduction*. Cambridge University Press.

The Georgian syllable in production and perception

Caroline Crouch, University of California Santa Barbara
crouch@ucsb.edu

The central question of this project is how the syllable is organized in space and time during production so that it is perceived as a cohesive unit. We focus on the syllable in Georgian specifically because of Georgian's incredibly permissive phonotactics: onsets can be up to seven consonants and there are no sonority-based restrictions. On the production side, we investigate the relationship between the sonority shape of a complex onset and a set of timing relationships within the onset. This involves a trio of Electromagnetic Articulography (EMA) experiments which reveal some of the ways in which time and space do and do not interact in shaping the Georgian syllable.

The first examines the relationship between sonority shape and local timing in the onset; the second, the relationship between sonority shape and global timing in the onset, and the third examines the effect of a morphological boundary on both global and local timing in the onset. We find a significant relationship between the local timing of consonant gestures in the onset and the sonority shape of that onset, but no relationship between sonority and global timing. Additionally, We confirm that the presence of a morphological boundary in the syllable onset does not affect timing relationships.

In the first experiment, two measures of gestural overlap were calculated: the interplateau interval, and relative overlap, which measures how early C2 begins relative to C1's target achievement. For both of these measures, sonority is significant. Sonority falls, like *rb* and *mt*, show earlier onset of C2, and shorter interplateau intervals. Rises, like *br* and *tm*, show later onset of C2, and longer interplateau intervals. Sonority plateaus, like *tb* and *mn*, are intermediate cases. We argue that long lag between gestures is the default timing pattern in Georgian, but is modulated in sonority falls in order to prevent intrusive vocoids from emerging in those onsets specifically, because they are more vulnerable to being misparsed as CVCV disyllables.

In the second experiment, we assess the presence of the c-center effect in Georgian but find no evidence for it. Instead, the data suggest anti-phase coordination between all gestures in Georgian, even between C and V in a CV syllable. This raises serious questions about the nature of the syllable in Georgian and presents a theoretical issue: Georgian syllables do not conform, either from an AP or a sonority-based perspective, to the structural expectations of a syllable, but are real units of language for speakers, as seen both in poetry and in native speaker judgments. We argue that this is possible because 1) there are no phonotactically illegal sequences in Georgian, so there is no need to distinguish between word-medial onsets and codas; and 2) Georgian prefers word-final open syllables, with codas often being the result of suffixation. Word-edge phonotactic probabilities lead speakers to interpret intervocalic sequences as maximally having a single coda consonant. This unique coordination pattern is motivated, we propose, in large part by the heavily prefixing verbal morphology of Georgian. The anti-phase only pattern supports easy 'slotting in' of these morphemes, as well as the perceptual recoverability of these critical grammatical elements.

In order to examine the relationship between syllable perception and gestural overlap that our results on C-C overlap suggest, we are currently designing a series of online perception experiments. Data are taken from the CC onset words from the production experiment, whose results are described above, and are selected in order to provide a representative range of overlap values. In these experiments we aim to test three hypotheses: 1) that a long constriction lag helps with C1 identification, 2) that short lag between C1 target and C2 onset helps preserve a CCV parse, and 3) that sonority falls are more vulnerable to this misparsing than sonority plateaus, which in turn are more vulnerable than sonority rises.

In these experiments, native Georgian speakers will be asked to perform two tasks. The first is to listen to audio clips and provide transcriptions in the Georgian orthography. Audio clips are taken from the production data to represent the full ranges of constriction lag and C1 target-C2 onset lag, found in the production data. Assessment of responses will be based on the transcriptions, which will be either correct or incorrect. This task will evaluate how well

constriction lag supports C1 identification. In some cases, transcriptions may also reveal CVCV misparsings, but given that all the data used represent felicitous productions, it is entirely possible that there will be few to no CVCV transcriptions.

We hypothesize that rises will have a high percentage of correct C1 IDs even at lower lags, since the less sonorous sound is being released into a more sonorous sound. For plateaus, correct C1 identification will increase as lag increases, because most C1s in the plateau data are stops. For falls correct C1 identification will likely be high regardless of lag due to segment-internal cues to identity in sonorants, which are the majority of C1s in the fall data. In general, there should be a low error rate in C1 identification.

In order to test hypotheses about syllabification and C1 target-C2 onset lag, participants will engage in an AXB forced choice task. They will hear two recordings of the first two consonants and vowel (CCV) of a Georgian word, each played twice in a row. Participants will then be asked which is a better start to the word. Lag values at either extreme of the spectrum should be judged to be less appropriate. We hypothesize that participants will be the most tolerant of a range of lag values in sonority rises, and least tolerant of extreme values for sonority falls. This reflects the stability of these measures as quantified by the interquartile range for each sonority shape as well.

The overall goal of this research project is to provide a definition of the syllable that addresses its spatial and temporal domains and is capable of capturing a wider range of syllables types, such as those in Georgian, which appear to defy current definitions of the syllable. Although some independent evidence for the syllable in Georgian does exist in prosodic research (e.g., Bush, 1999; Skopeteas and Féry, 2016) and in contemporary poetry, where syllable counts are used for haiku, this work aims to provide evidence from articulatory and perceptual data to understand what spatial and temporal aspects of the syllable are necessary for it to function as a unit in Georgian.

References

Bush, Ryan (1999). Georgian yes–no question intonation. Phonology at Santa Cruz, Vol. 6. UC Santa Cruz, Santa Cruz, CA. 1–11.

Crouch, C., Katsika, A., Chitoran, I. 2020. The role of sonority profile and order of place of articulation on gestural overlap in Georgian. Proc. 10th International Conference on Speech Prosody 2020, (pp. 205-209) DOI: 10.21437/SpeechProsody.2020-42.

Nam, Hosung, Louis Goldstein, and Elliot Saltzman. "Self-organization of syllable structure: A coupled oscillator model." Approaches to phonological complexity 16 (2009): 299-328.

Skopeteas, S., & Féry, C. (2016). Focus and intonation in Georgian: Constituent structure and prosodic realization. Manuscript.

Uncertainty does not affect phonetic convergence in a scripted dialogue

Dorina de Jong, University of Ferrara & Istituto Italiano di Tecnologia
dorina.dejong@iit.it

Phonetic convergence is the observation that conversational partners will often begin to sound more alike as their conversation progresses (Pardo, 2006). According to a prominent theory, this phenomenon relies on forward and inverse models where one makes predictions of one's interlocutor's articulation. When one's prediction is not the same as the actual articulation of one's interlocutor, the following prediction errors lead one to adjust their speech production system for better subsequent simulations and predictions (Pickering and Garrod, 2013; Gambi and Pickering, 2013). But we know little about if and how phonetic convergence might change as a function of the predictability of their interlocutor.

The main objective of our current study is to see whether we can alter the degree of phonetic convergence through violating predictions in a scripted dialogue. To be more specific, we let nine Italian-native (21 to 33 years old, average age 24.50 ± 3.65; 3 male) and ten French-native (19 to 42 years old, average age 23.45 ± 4.94; all female) same-sex dyads alternately read aloud a neutral text in English for 80 speaking turns, known as an alternating reading task (see Aubenel and Nguyen, 2020), and violated predictions by creating discrepancies between what the listening participant would expect and what they would hear their interlocutor say. The listener was able to create expectations on what the speaker would say, because the listener could read along silently with the speaker. The scripted dialogue allowed us to introduce violations of predictions in a controlled manner where the speaker would say a synonym instead of the word that the listener saw on their screen. The word that the listener would hear was, therefore, semantically correct but unexpected. Participants read the same text together for four times, varying the shown text's uncertainty by presenting synonyms in 0%, 25%, 50% or 75% of the speaking turns. Phonetic convergence was extracted using an automatic speaker identification technique based on Gaussian Mixture Model - Universal Background Model (GMM-UBM, see Mukherjee et al., 2019 for the model on word-level). Features assessed were Mel-frequency cepstral coefficients to get a broad view of the phonetic features of our participants. The GMM-UBM gives two outcomes: the log-likelihood ratio (LLR) quantifies the similarity in one's speech to the interlocutor, with maximal similarity at zero, while one could also set a threshold of phonetic convergence.

We expected to see greater phonetic convergence when dyads interacted within an uncertain situation. For one, because we thought that the increase in uncertainty would elevate the participants' attention, thereby possibly freeing more cognitive resources to the benefit of phonetic convergence. Moreover, an increase in uncertainty would also introduce more cues (= prediction errors) to update the speech production system of the listener. On the other hand, the increase in uncertainty could also instead occupy attentional resources, thereby hindering phonetic convergence. Furthermore, unpredictable situations also constitute misalignment in the conversation, which could lead to phonetic misalignment. After all, levels of alignment are thought to be interconnected (Pickering and Garrod, 2004).

However, results from a linear mixed-effects model (LMEM) with condition as dependent variable, subject as random slope and subject, order and sentence number as random intercept show that the LLR score of "0%" does not differ from "25%" (β = -.01, SE = .04, p=.83), "50%" (β = .04, SE = .05, p=.40), nor "75%" (β = -.04. SE = .04, p=.32). The result does not differ between the Italian and the French natives (p=.56) and is not dependent on the participants' English proficiency (p=.14, assessed by LexTALE, Lemhöfer and Broersma, 2012). Results from a binomial generalized linear mixed model (GLMM) show that it is also not more probable to see more or less phonetic convergence in "25%" (odds ratios = .96, 95% CI [.68-1.37], z = -.20, p = .84), "50%" (.85 [.58-1.24], z = -0.86, p = .39), and "75%" (1.08 [.76 - 1.54], z = .44, p = .66) when compared to the "0%" condition. Again, the results are independent from the native language of the participant (p=.22) and their English proficiency (p=.07). Together, these results indicate that uncertainty does not affect the observed phonetic convergence in a scripted dialogue. However, it could be that the method used was not strong enough to create differences.

References

Aubanel, V., & Nguyen, N. (2020). Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task. *PloS one*, 15(5), e0232209.

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in psychology,* 4, 340.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior research methods*, 44(2), 325-343.

Mukherjee, S., Badino, L., Hilt, P. M., Tomassini, A., Inuggi, A., Fadiga, L., ... & d'Ausilio, A. (2019). The neural oscillatory markers of phonetic convergence during verbal interaction. *Human brain mapping*, 40(1), 187-201.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4), 2382-2393.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2), 169-190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences,* 36(4), 329-347.

Articulatory overlap as a function of stiffness in German, English and Spanish word-initial stop-lateral clusters

Shihao Du, Universität Potsdam
shihao.du@uni-potsdam.de

In the past three decades a number of parameters have been identified to affect inter-segmental overlap in CC clusters including cluster position, place order, C1 voicing and place, C2 manner, vocalic context, speech rate and prosody etc. But little attention has been paid to the parameters of the underlying dynamical system assumed to generate the differing degrees of overlap in the articulatory gestures comprising these CC clusters.

Browman and Goldstein (1990) first pointed to the possibility that different degrees of overlap may be related to some notion of rapidity of the different oral articulators implicated in the consonants of the cluster. They suggested that some gestures are more easily overlapped or “hidden” by other gestures due to differences in rapidity. Specifically, they postulated that a slower gesture, such as that implicated in a labial or a velar constriction, “might prove more difficult to hide” than a faster one as in a coronal, since the tongue tip enjoys a greater flexibility than the lips or more posterior parts of the tongue such as the tongue body implicated in velars (Browman & Goldstein, 1990, p. 18).

The suggested link between overlap and some notion of rapidity by Browman and Goldstein (1990) was pursued systematically by Jun (2004). Jun proposes that in a C1C2 consonant cluster, when the rapidity of C2 is held constant, a more rapid C1 articulatory movement would result in more overlap between the gestures of C1 and C2 than a slower C1. Conversely, when the rapidity of C1 is held constant, a slower C2 would lead to more articulatory overlap between the gestures of C1 and C2 than a faster, as the C2 gesture would “intrude” into the unfolding of the C1 gesture. Recently, Roon, Hoole, Zeroual, Du, Gafos (2021) tested Jun’s proposal about the relationship between rapidity and overlap using electromagnetic articulatory data from Moroccan Arabic word-medial stop-stop clusters. Roon et al. (2021) found no systematic differences in rapidity and the predicted ordering of overlap based on articulator, regardless of whether rapidity was indexed by peak velocity or stiffness. But they did find a strong correlation between overlap and the difference between C1 and C2 stiffness.

The present study extends this line of investigation in two ways. First, it tests the findings from one language as in Roon et al. (2021) against articulatory data from three more languages that showed distinct temporal organization patterns, namely word-initial stop-lateral clusters in German, English and Spanish. Second, it disentangles the individual contributions of C1 and C2 stiffness to overlap, an issue which was not studied explicitly in Roon et al. (2021). In the current work, former results on inherent velocities and stiffness difference were successfully replicated for all three languages. However, it was also found that the stiffness of the C2 closing gesture had a more robust effect on overlap than that of C1, as shown in **Figure 1**. The present study concludes that overlap in word-initial C1C2 clusters is primarily determined by the control parameter of stiffness of the second consonant as opposed to that of the first consonant. This result is placed in the context of other work on articulatory and perceptual influences on inter-consonantal timing as well as in the broader context of work on the coupling between perception and action in other domains of skilled action.

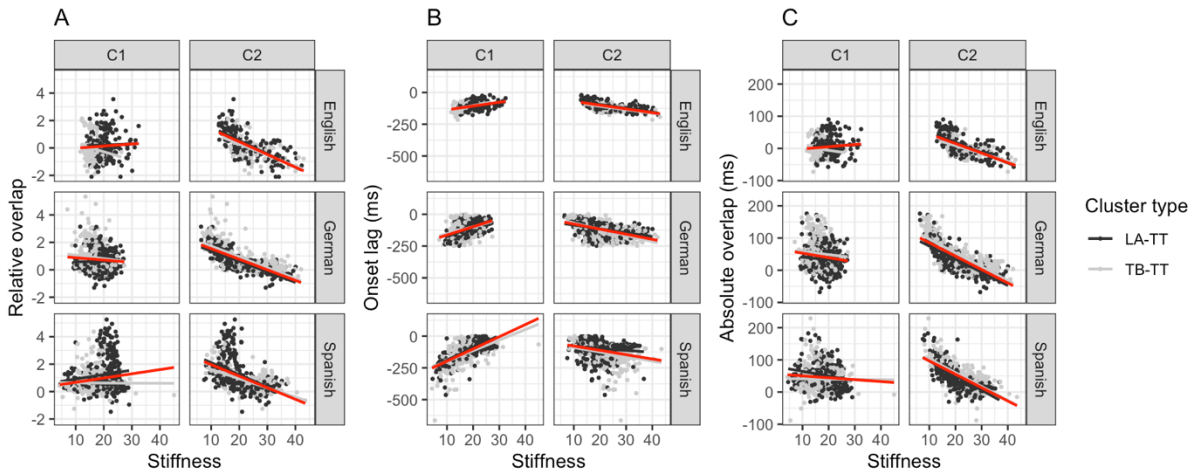


Figure 1: Scatterplots showing relations between the stiffness of C1/C2 closing movement and three temporal overlap measures: relative overlap (A), onset lag (B), and absolute overlap (C) across cluster types and languages. Compared to C1 stiffness, C2 stiffness has a more robust relation with the different overlap measures.

References

Browman, C. P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. E. Beckman (Eds.), Papers in Laboratory Phonology (1st ed., pp. 341–376). Cambridge University Press. https://doi.org/10.1017/CBO9780511627736.019

Jun, J. (2004). Place assimilation. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), Phonetically Based Phonology (pp. 58–86). Cambridge University Press.

Roon, K. D., Hoole, P., Zeroual, C., Du, S., & Gafos, A. I. (2021). Stiffness and articulatory overlap in Moroccan Arabic consonant clusters. Laboratory Phonology: Journal of the Association for Laboratory Phonology, 12(1), 8. https://doi.org/10.5334/labphon.272

The perception-production link in prosody – investigated by prosodic cue use for disambiguation of structural boundaries

Andrea Hofmann, Kognitionswissenschaften, Department Linguistik, Universität Potsdam
andrea.hofmann@uni-potsdam.de

The presented dissertation project draft is part of research project B01 of the SFB1287 and focuses on different perspectives of prosodic cue use in perception, production, and their potential interaction. We build on our previous findings where we identified individual differences in production and perception experiments: Firstly, in perception, a subgroup of participants was specifically sensitive to prosody using subtle early prosodic cues to reliably predict the intended syntactic structure in sentences with/without an internal phrase boundary like (1) and (2), while the other subgroup made their decision later, after all prosodic cues were processed (Hansen et al., submitted). Secondly, in production, a subgroup of participants produced all of the three main cues used for prosodic boundary marking (lengthening, pitch rise, pause) distinctly to distinguish between sentences like (1) and (2), while the others only used two or one of the cues (Huttenlauch et al., 2021).

Our goal is to investigate both modalities within the same participants which enables us to test the coupling between a speaker's perceptual acuity and distinctness of production in the domain of prosody. A perception-production link is predicted through prosodic control processes in the gradient order DIVA (GODIVA) model (Bohland et al., 2010). We strive to answer to what extent individual differences in prosodic boundary cue production systematically couple with their perceptual sensitivity and whether this coupling is modulated by individual differences in general cognitive abilities.

For this purpose, we are compiling a series of experimental tasks and behavioral measures as a comprehensive overview of individual performances which shall be collected in 50 healthy young German adults. Participants ought to produce and perceive stimuli where the same sequence of lexical items can have different meanings dependent on whether or not a prosodic phrase boundary is inserted. The intended meaning is signified as an internal grouping of coordinated name (Kentner & Féry, 2013) or noun (Zhang, 2012) sequences:

(1) sequence with internal grouping (bracket)	(2) sequence without internal grouping (no-bracket)
"(Moni und Nelli) und Lola"	"Moni und Nelli und Lola"
(single nouns)	(composite noun)
“Thunfisch, Salat und Cola”	“Thunfischsalat und Cola”

We will investigate the expression of three main prosodic cues found in German to prosodically mark structural boundaries: (a) final lengthening: the lengthening of the pre-boundary syllables, (b) f0-range: an increase in fundamental frequency (f0) of the pre-boundary syllables, and (c) pause duration: the insertion or prolongation of a pause at the prosodic boundary (after "Nelli" or “Thunfisch” in example (1) (Kentner & Féry, 2013; Huttenlauch et al., 2021).

We expect all participants to differentiate between conditions (1) and (2) in production and perception. Specifically, we hypothesize that individuals who produce all prosodic cues (a, b, c) distinctively (possibly even as early as on the first name / noun) will be the same individuals who are most sensitive to prosodic cues in perception and who, for instance, can use subtle differences in early prosodic cues for prediction of the upcoming structure. To arrive at a comprehensive overview on individual patterns and behavior we plan to test our hypotheses through a variety of experimental paradigms.

Perception & Production: We will start with a joint examination of individual differences in perception and production by means of the magnitude of auditory sensorimotor adaptation. We will use a similar procedure to compare the results with individual discrimination abilities. In addition, we investigate if these individual patterns are reflected in the ability to imitate fine details from a speech model. And we will complete the overview with a forced-choice task to investigate differences in the sensitivity to the amount of prosodic information available.

We expect that individuals that are most adaptive to perturbed auditory feedback will show better discrimination skills and imitate model stimuli more closely. Furthermore, we expect the same individuals to consistently choose the correct grouping condition early on.

- Perturbation paradigm: Participants' oral responses (to (1) and (2)) are recorded, manipulated, and played back as "perturbed" auditory feedback in real time. This is preceded by a block with unperturbed productions to assess individual baseline values for each cue. If individuals show compensatory / adaptive responses when a manipulated f0 and syllable duration of their own voice is played back to them, this would hint to a coupling of the speech motor control system and speech perception (Bohland et al., 2010).

- Discrimination task: The above procedure will be combined with an active and a passive discrimination task (just noticeable difference (JND) procedure, Villacorta et al., 2007; Lester-Smith et al., 2020).

- Imitation task: Participants have to repeat pre-recorded versions of the stimuli as similar as possible. The recordings stem from a female speaker in which the salience of the cues has been manipulated (De Beer et al., submitted). We will calculate similarity for all cues.

- Gating paradigm: Participants listen to a controlled set of sequences with/without internal grouping (see (1) and (2)) presented in syllable length snippets / gates ascending from one to all syllables of the sequence and have to decide about the intended grouping after each gate (Hansen et al., submitted).

Cognitive tests: To determine possible mediating factors for the behavioral patterns we discover, we will administer and later correlate a range of tests targeting cognitive functions for which we presume an influence on the use of linguistic prosody (pragmatic and socio-cognitive skills, working memory tasks, acoustic abilities, general processing speed, auditory-motor synchronization, e.g. Assaneo et al., 2019).

We will model all collected measures as predictors in Bayesian mixed-effects multivariate regression testing for individual differences in the coupling of perception and production.

References

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). *Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning*. Nature Neuroscience, 22(4), 627-632.

Bohland, J. W., Bullock, D., & Guenther, F. H. (2010). *Neural representations and mechanisms for the performance of simple speech sequences*. Journal of Cognitive Neuroscience, 22(7), 1504-1529.

De Beer, C., Hofmann, A., Regenbrecht, F., Huttenlauch, C., Wartenburger, I., Obrig, H. & Hanne, S. [Manuscript submitted for publication]. *Production and comprehension of prosodic boundary marking in persons with unilateral brain lesions*.

Hansen, M., Huttenlauch, C., de Beer, C., Wartenburger, I. & Hanne, S. [Manuscript submitted for publication]. *Individual differences in early disambiguation of prosodic cues*.

Huttenlauch, C., de Beer, C., Hanne, S. & Wartenburger, I. (2021). *Production of prosodic cues in coordinate name sequences addressing varying interlocutors*. Journal of the Association for Laboratory Phonology, 12(1).

Kentner, G. & Féry, C. (2013). *A new approach to prosodic grouping*. The Linguistic Review, 30(2), 277–311.

Lester-Smith, R. A., Daliri, A., Enos, N., Abur, D., Lupiani, A. A., Letcher, S., & Stepp, C. E. (2020). *The relation of articulatory and vocal auditory–motor control in typical speakers*. Journal of Speech, Language, and Hearing Research, 63(11), 3628-3642.

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). *Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception*. The Journal of the Acoustical Society of America, 122(4), 2306-2319.

Zhang, X. (2012). *A comparison of cue-weighting in the perception of prosodic phrase boundaries in English and Chinese* (Doctoral dissertation, University of Michigan).

Exploring the link between speech sound perception and production

Lena-Marie Huttner, Laboratoire Parole et Langage Aix-Marseille Université
Lena-marie.huttner@univ-amu.fr

Over the course of an interaction interlocutors tend to become more similar in linguistic behavior (Pickering & Garrod, 2004). In the domain of speech, this phenomenon is referred to as phonetic convergence (Natale, 1975). Interlocutors' acoustic-phonetic realization will tend to become more similar over the course of an interaction. Similarly, research on perceptual learning has found, that people will tend to adjust the perception of sounds to the input, i.e. their interlocutor (Liebermann, 1957; Kraljic & Samuel, 2005).

According to sensorimotor theories of speech (Schwarz et al., 2012) these two phenomena are not separate, but two sides of the same coin. The underlying cause of phonetic convergence is often hypothesized to be an inherent link between perception and production: in order for people to produce speech more similarly to one another, they must first perceive the sounds and then fine tune their production to the perceived input (Nguyen & Delvaux, 2015; Sato et al. 2013). However, exploration of this theory in the speech sound domain is sparse. We venture to explore this link experimentally by combining experiments in phonetic convergence and perceptual learning in a single study, by asking the following question:

Is a change in acoustic phonetic realization elicited by short-term social interaction accompanied by a change in perception?

Phonetic convergence has been elicited on several phonetic acoustic features, and is often assessed holistically through AXB perceptual tasks, whereas perceptual learning experiments create sound continua of one specific acoustic feature. To combine the two experimental approaches we will use stimuli on a 9-step VOT continuum, more specifically /p/ and /b/ spoken by an English native speaker and altered using a Praat script (Boersma & Weenik, 2022; Winn, 2020).

In a pre-post design we assess participant's perception of the stimuli as well as the production of target words ("bug" and "pug") before and after playing a picture naming game with a pre-recorded speaker.

Participants will be asked to produce the English words "bug" and "pug" and will then perform a categorization task to assess their a priori categorical boundary.

Participants will then perform a picture naming game with a pre-recorded speaker, believing it was another human in a different room. Participants and the pre-recorded speaker will take turns naming a categorizing the stimuli. After each turn, participants will receive feedback on whether they chose the response the speaker produced. The speaker will be biased towards on side of the continuum. Participants will then themselves name on of the pictures, for the imagined partner to guess.

Afterwards, participants will once again produce the target words.

Participants will be grouped into three conditions: In one, the target words will be produced during the game in another, the pre-recorded voice will produce the target sounds, but participants will produce fricatives. In a third condition participants will not produce the sounds at all, but just perform the naming task.

The data will then be analyzed to explore if: there was a change in production, whether the categorization differs between the pre and post categorization tasks, and if the change in perception can be predicted by the bias of the stimuli in the picture naming game, or by the degree of convergence, i.e. the participant's own production.

We will further explore whether a possible convergence/perceptual learning effect is specific to the lexical items used in the previous task, or whether it translates to other lexical items as well. Further, we will be able to explore whether this effect translates to other voices, or is speaker-specific.

This experiment can be seen as a stepping stone towards exploring the perception-production link in speech sound categorization.

References

Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.14, retrieved 24 May 2022 from http://www.praat.org/

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. Cognitive psychology, 51(2), 141-178.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of experimental psychology, 54(5), 358.

Natale, M. (1975). Social desirability as related to convergence of temporal speech patterns. Perceptual and Motor Skills, 40(3), 827-830.

Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. Journal of Phonetics, 53, 46-54.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(2), 169-190.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. Frontiers in Psychology, 4.

Schwartz, J. L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5), 336-354.

Automatic detection of nasality for speaker characterisation

Lila KIM, Laboratoire de Phonétique et Phonologie
lila.kim@sorbonne-nouvelle.fr

There are very few languages where nasality is not present phonologically: 96% of the world's languages have at least one nasal consonant. Nasal vowels can also be present in languages for which they are not phonemes, such as English through the notion of nasal coarticulation.

Nasality is achieved by the coupling of two cavities, nasal and oral, by lowering of the velum, which produces acoustic effects on nasal sounds (Havel, 2016) such as the introduction of nasal resonances, the reduction of energy in the spectrum, changes to overall formant structures and to the overall spectral envelope of the vowels (Styler, 2017). These changes make acoustic analyses difficult and complex, whether manual or automatic, because they are very sensitive to various variations such as accentuation, sound articulation, language, etc. Acoustic analyses are made even more difficult because the lowering of the soft palate can also result in altering the shape of the oral cavity and thus altering the formants (Carignan, 2017).

Voice quality is considered in the literature to have great implications for the speaker characterization (Gold & French, 2019). It can be a permanent part of a speaker's voice due to physiological peculiarities (mainly laryngeal and supralaryngeal) or speaker's habits (sociolinguistic factors), but it is also subject to intra-speaker variability, especially in the style of speech or emotion (Nolan, 2005). If the voice quality does not identify a speaker, it does allow a reliable characteristic to be provided in addition to other characteristics such as the fundamental frequency or the articulation of vowels and consonants. In the field of automatic speaker recognition, it has frequently been observed (Kahn et al., 2011) that nasals are more relevant in characterizing the speaker than other phonemes. The explanation given is that the nasal cavity is both different from one speaker to another, and also not very malleable during the production of speech (Dang et al., 1994; Serrurier, 2006), which would make it a stable resonance cavity specific to the speaker.

Table 1: Confusion matrix for nasal vs. non-nasal classification results obtained by cnn

	F-score nasal	**F-score non-nasal**	**Accuracy**
/a/		96%	94%
/ɑ̃/	97%		93%

It was tried to evaluate with Convolutional Neural Networks (CNN), the ability to discriminate acoustically an oral vowel from a nasal vowel (especilly with vowels /a/ and /ɑ̃/) on a corpus of 45 French-speaking speakers, as well as the possibility to generalize this discrimination on speakers and/or vowels not previously trained. For a nasal vowel vs. oral vowel classification, the results obtained could reach up to 94% of correct identification. Also, there was work on the effect of some factors in the malclassification obtained by the networks, which were a speaker, the duration of the vowel and phonemic contexts of the vowel. It was observed that oral vowels followed by a pause or a dorsal consonant were classified as nasal according to our neural network. Some speakers lower the soft palate at the end of a sentence by relaxation or rest, which may result in a nasalized tone, and explain why the network incorrectly classifies oral vowels in this context.

The goal of future research will be to improve the results by training the neural network with several hundred hours of speech to recognize nasality in all phonemes and apply them to a larger population. In order to validate the results, a physiological measurement will be used to obtain a nasal airflow over time, a verification corpus will be created thanks to the aerodynamic method to be able to compare the nasal airflow with the nasality probabilities obtained by our models. Finally, since the CNN models use images, the masking technique

will be discussed to visualize the areas where the interesting and used information is located for their decision making.

References

Carinan, C. (2021). A practical method of estimating the time-varying degree of vowel nasalization from acoustic features. *The Journal of the Acoustical Society of America*, 149(2), 911-922.

Dang, J., Honda, K., & Suzuki, H. (1994). Morphological and acoustical analysis of the nasal and the paranasal cavities. *The Journal of the Acoustical Society of America,* 96(4), 2088-2100.

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

Havel, M., Kornes, T., Weitzberg, E., Jon, O., & Sundberg, L. & J. (2016). Eliminating paranasal sinus resonance and its effects on acoustic properties of the nasal tract. *Logopedics Phoniatrics Vocology,* 41(1), 33-40.

Nolan, F. (2005). Forensic Speaker Identification and the Phonetic. *A Figure of Speech: A Festschrift for John Laver*, 385.

Kahn, J., Audibert, N., Bonastre, J.-F., & Rossato, S. (2011). Inter and Intra-speaker Variability in French: An Analysis of Oral Vowels and Its Implication for Automatic Speaker Verification. *ICPhS*, 1002-1005.

Serrurier, A. (2006). Modélisation tridimensionnelle des organes de la parole à partir d'images IRM pour la production de nasales - Caractérisation articulatori-acoustique des mouvements du voile du palais, *PhD Thesis*. INP Grenoble.

Styler, W. (2017). On the acoustical features of vowel nasality in English and French. *The Journal of the Acoustical Society of America*, 142(4), 2469-2482.

Comparison of methods used to quantify acoustic-prosodic entrainment

Joanna Kruyt, Institute of Informatics, Slovak Academy of Sciences / Faculty of Informatics and Information Technology, Slovak Technical University
joanna.kruyt@savba.sk

Abstract: Comparison of methods used to quantify acoustic-prosodic entrainment. During an interaction, people tend to behave more similarly. Such increased similarity in behaviour is often referred to as “entrainment”, though a large number of other terms exist, such as accommodation, alignment, convergence, et cetera. Perhaps the most well-researched behaviour in which it occurs is language: entrainment has been observed at practically all levels of language, including syntax (e.g. Branigan, Pickering & Cleland, 2000), lexical choice (e.g. Brennan & Clark, 1996; Garrod & Anderson, 1987), and prosody (e.g. Levitan et al., 2012; Natale, 1975). Entrainment is associated with positive social measures such as effective and satisfying communication (e.g. Chartrand & Bargh, 1999). It has been hypothesised that entrainment is not one single latent behaviour, but that entrainment on different features, levels, and dimensions may be governed by different mechanisms or serve different purposes in interaction (Weise & Levitan, 2018; Ostrand & Chodroff, 2021).

Importantly, a vast range of methods exist to determine or quantify entrainment. This is especially true for acoustic-prosodic entrainment, which has been researched by linguists, psychologists, and computer scientists, who each brought their own methods to the table. There is little consistency in entrainment research in terms of terminology and methodology, despite attempts to ameliorate this. The present study provides an overview and discussion of common, influential, or conceptually interesting methods used to determine acoustic-prosodic entrainment.

Additionally, a total of 11 methods were compared: the same 3 features (mean fundamental frequency, fundamental frequency range, maximum fundamental frequency) from the same conversations were analysed with these 11 methods, to see whether they provide the same results, or whether they result in different findings and perhaps measure different aspects or dimensions of entrainment. The 11 methods from the following papers were compared: global proximity, global convergence, local proximity, local convergence, local synchrony (Levitan & Hirshberg, 2011), prediction using linear mixed effects models (Schweitzer & Lewadowski, 2013), geometric approach (Lehnert-LeHouillier, Terrazas, & Sandoval, 2020), time-aligned moving average (Kousidis et al., 2008), prosodic accommodation dynamics tool (De Looze et al., 2014), cross-recurrence quantification analysis (e.g. Fusaroli & Tylén, 2016), and windowed-lagged cross-correlation (Boker et al., 2002). The methods employed in this study differed in various ways: for example, some measure acoustic-prosodic entrainment on the turn-level, while others measure it over the whole conversation. Results from the different methods will be presented on a poster.

Results show that the methods indeed capture different aspects, types, or dimensions of entrainment, which provides further support to the notion that entrainment is not one behaviour mediated by the same mechanisms on all levels, but rather is highly complex and influenced by many different factors. In short, the goal of the present study was three-fold: it aimed to a) provide an overview of commonly used methods in acoustic-prosodic entrainment research and highlight the terminological and methodological inconsistencies in the field, b) illustrate that different methods capture the same phenomenon in the same way, suggesting that “acoustic-prosodic entrainment” perhaps is not one single behaviour but consists of several subtypes or -behaviours, and c) to make the code for the analyses publicly available,, to facilitate consistency in methodologies and to make acoustic-prosodic entrainment research more widely accessible.

References

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological methods, 7(3), 338.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. Cognition, 75(2), B13-B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22(6), 1482.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. Journal of personality and social psychology, 76(6), 893.

De Looze, C., Scherer, S., Vaughan, B., & Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. Speech Communication, 58, 11-34.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. Cognitive science, 40(1), 145-171.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition, 27(2), 181-218.

Kousidis, S., Dorran, D., Wang, Y., Vaughan, B., Cullen, C., Campbell, D., ... & Coyle, E. (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. In Ninth Annual Conference of the International Speech Communication Association.

Lehnert-LeHouillier, H., Terrazas, S., & Sandoval, S. (2020). Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum. Frontiers in Psychology, 11, 2718.

Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 11-19).

Levitan, R., & Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In Twelfth Annual Conference of the International Speech Communication Association.

Natale, M. (1975). Social desirability as related to convergence of temporal speech patterns. Perceptual and Motor Skills, 40(3), 827-830.

Ostrand, R., & Chodroff, E. (2021). It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. Journal of phonetics, 88, 101074.

Schweitzer, A., & Lewandowski, N. (2013, August). Convergence of articulation rate in spontaneous speech. In INTERSPEECH (pp. 525-529).

Weise, A., & Levitan, R. (2018, June). Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 297-302).

200 Greek Idiomatic Expressions: Ratings for Familiarity, Ambiguity and Decomposability

Anastasia Lada[1], Philippe Paquier[1], Ifigeneia Dosi[2], Christina Manouilidou[3], Stefanie Keulen[1]

Vrije Universiteit Brussel[1], Democritus University of Thrace[2], University of Ljubljana[3]

Anastasia.Lada@vub.be, Philippe.Paquier@vub.be, idosi@helit.duth.gr, christina.manouilidou@ff.uni-lj.si, Stefanie.Keulen@vub.be

Introduction

Idioms differ from other forms of figuration because of their semantic dimensions of familiarity (frequency of encounter), ambiguity (possibility to have a literal interpretation) and decomposability (possibility of the idiom's words to assist in its figurative interpretation) (Langlotz, 2006). This study focuses on the Greek language and seeks to provide further insights into idiom processing. Research in Greek is limited even though Greek idioms bear some very distinct characteristics that make them a good candidate to explore idiom processing: they can be categorised based on all semantic dimensions (Vlaxopoulos, 2007), they have extremely high productivity and ability for compositionality (Papalexandrou, 2014) as for example their constituent words can even be fake words or non-words (Mazi, 2012) and they have great syntactic flexibility as for example they can even appear as whole sentences (Vlaxopoulos, 2007).

Aims

This study aimed at providing a corpus of 200 Greek idioms rated by 50 native Greek raters, aged 20-38 years (M= 22,6; SD=4,023), 40 females and 10 males, all postgraduate students at Democritus University of Thrace (Greece). Specifically, the study aimed at (1) rating all idioms in terms of their degree of familiarity, ambiguity, and decomposability and (2) investigating the associations among these dimensions, providing the first corpus of Greek idioms rated for semantic dimensions.

Methods

We conducted 3 different online assessments each of which asked the participants to evaluate the degree of an idiom's familiarity, ambiguity, and decomposability. The idioms were selected from the dictionary of Greek Idioms: "Dictionary of Idioms in Modern Greek" (Vlaxopoulos, 2007). Each list had the same 200 idioms. Participants were asked to rate idioms on a Likert scale, ranging from 0 to 5 (0 corresponding to low and 5 high degree of ambiguity, familiarity, and decomposability). Methods followed the studies of Libben and Titone (2008) and Titone and Connine (1994b.)

Results

Cronbach's alpha and Intraclass Correlation Coefficient (Hubers et al., 2019) verified high internal consistency in the data. Familiarity was positively correlated with decomposability ($rs=.409$, $p<.01$) and weakly with ambiguity ($rs=.189$, $p<.01$). Last, Mann Whitney U tests demonstrated that familiarity ratings showed significant differences with decomposability ratings ($U=5.510$, $p=0.002$). Therefore, considering that the most frequent idioms in a language are the highly familiar, then it is obvious that Greek idioms are decomposable in their majority. Decomposable idioms have constituent words that are linked to the figurative meaning. This fact would well explain the high semantic productivity that exists in Greek idioms. The constituent words could be for example replaced with other semantically related words and we would come up with alternative idioms which would still be semantically and pragmatically equal (Mazi, 2012). In such cases, the speakers would be still able to recognize the idiom. Also, in the Greek language, the more familiar idioms tend to be more ambiguous in agreement with Vlaxopoulos (2007) hypothesis. Last, the results allow for the creation of a corpus with idioms rated in all their semantic dimensions and facilitate future research in Greek.

References

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative data of Dutch idiomatic expressions: Subjective judgments you can bank on. Frontiers in psychology, 10, 1075.

Langlotz, A. (2006). Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English (Vol. 17). John Benjamins Publishing.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. Memory & Cognition, 36(6), 1103-1121.

Mazi, V. D. (2014), The use of journalistic texts in the teaching of Greek as a foreign language: the case of privatizations (No. GRI-2014-13065). Aristotle University of Thessaloniki.

Papalexandrou, E. (2014). Comparative approach of Russian and Greek phrasalisms: references to excerpts from N. Gogol "Diary of a madman" and N. Kazantzakis "The last temptation".

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. Metaphor and Symbol, 9(4), 247-270.

Vlaxopoulos, S., (2007). Dictionary of Idioms in Modern Greek. Kleidarithmos, ISBN: 9789604610075 (In Greek)

Preliminary acoustic study: Phonetic divergence in simultaneous bilinguals, language learners, and monolinguals

Andres Lara, LPP, CNRS UMR-7018 / Sorbonne Nouvelle University
andres.lara@sorbonne-nouvelle.fr

This preliminary study measured divergence at the segmental level in English and in French. We compared the shadowed productions of American language learners of French and simultaneous bilinguals to those of monolingual speakers. Language learners were expected to deviate in French (L2) but not in English (L1) due to interlinguistic transfer (Flege, 1995). According to Paradis (2001) simultaneous bilinguals have two separate phonological systems that exert mutual influence on each other as stated in the Interactional Dual Systems Model. We can then expect bidirectional influence in bilingual productions causing them to deviate from the stimuli values (TV) in both languages. Last but not least, we questioned whether convergence with TV correlates with phonological competence in a foreign language.

F1 to F3 formants values of American English (AE) /i ɪ ɛ æ ɑ ɔ ʊ u/ and French /i e ɛ y ø œ a ɔ o u/ vowels were analyzed for 5 simultaneous bilinguals (BL, age 27, SD = 4) living in France having been exposed to both languages from birth, 5 American learners of French (LL: age 24 SD = 2) exposed after the age of 13 and having lived in France for 3+ years and, 10 monolinguals (ML: age 30, SD = 7) living either in France (5) or in the USA (5). Vowels were recorded in three phonetic contexts: isolated, /bVb/, and /pVp/. Formants were extracted at midpoint along with duration values for all vowels. The study included 15 women and 5 men; all formant values were normalized using Labov's method (Labov, 2006). A Shapiro Wilk test revealed abnormal distributions for English and French. A Kruskal Wallis (McKnight & Najab, 2010) test was used as an alternative to the ANOVA test.

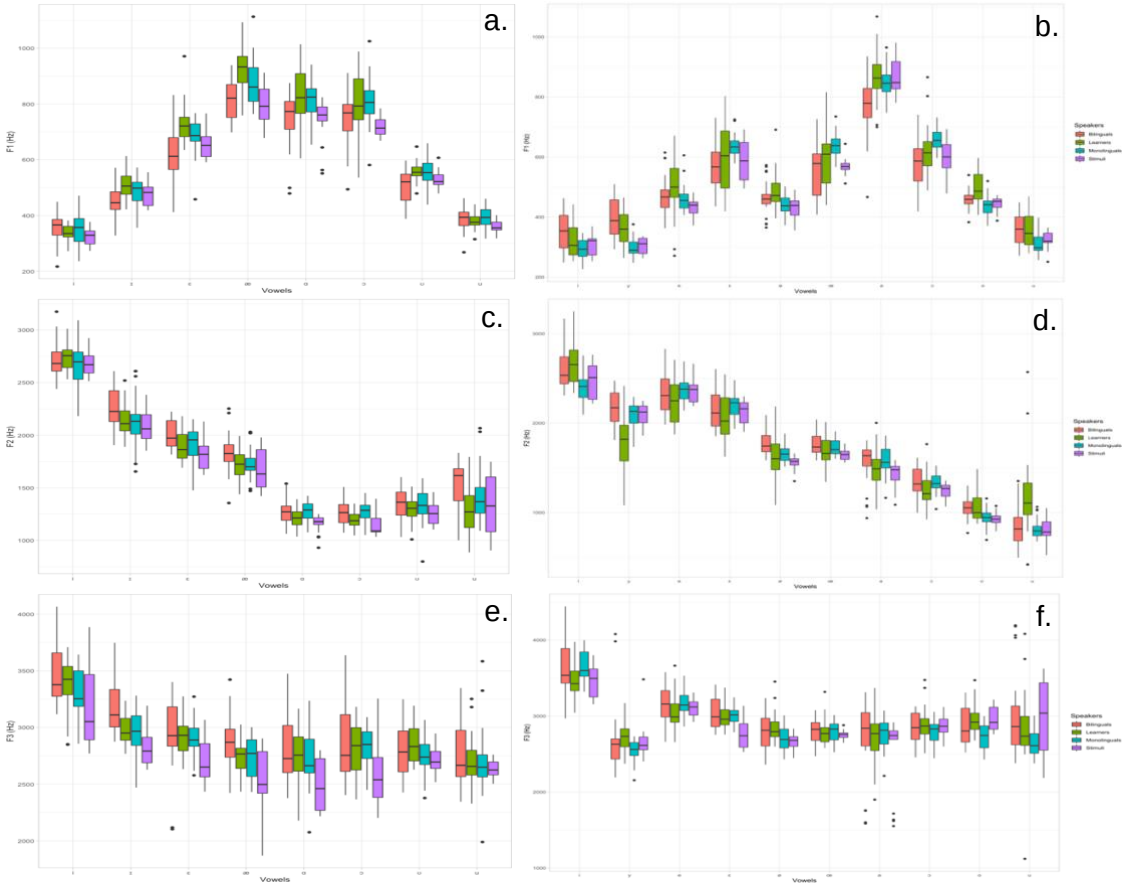


Figure 1. F1 (a), F2 (c) and F3 (e) values for English vowels.
F1 (b), F2 (d) and F3 (e) values French vowels.

Significant differences between speaker profiles (SP) for English F1 values ($N_{(3, 867)}$ = 12.364; p > .006). Language learners deviated significantly from TV. BL and ML did not deviate from TV. LL and ML produced open-mid /ɛ æ ɔ/ and open /ɑ/ vowels more open than those modeled by TV (Fig. 1a). Significant divergences between SP ($N_{(3, 867)}$ = 9.988; p > .018) for F2 values. BL deviated significantly from TV; F2 was higher for BL (see high [ɪ u], mid [ɛ], and low [æ] vowels in Fig. 1c). Significant divergences ($N_{(3, 867)}$ = 56.775; p < .0002) for F3 values. All SP deviated significantly from TV. F3 was higher for all SP (Fig. 1e); especially for high [i ɪ u], mid [ɛ ɔ], and low [æ ɑ] vowels in BL; mid [ɛ ɔ] and low [æ ɑ] in the other speakers. Significant differences in duration between the SP ($N_{(3, 867)}$ = 12.184; p < .007). BL and ML deviated with shorter vowels as a whole. The intrinsic duration-related distinction between tense and lax vowels was well maintained by all speakers.

Comparison of French F1 values did not reveal any significant differences ($N_{(3, 1082)}$ = 7.32; p < .062). ML deviated the least out of all SP. BL and LL deviated for the vowel contrast /u – y/ (Fig. 1b). F1 for /y/ was higher for BL (392Hz) and LL (361Hz) than it was for both ML (298Hz) and TV (304Hz). F1 values for low vowel [a] in BL (773Hz) were lower than TV (864Hz). LL F2 values for vowel contrast /u – y/ show differentiation problems. A reduced formantic space between /u – y/ due to higher values for [u] and lower values for [y] (Fig. 1d). LL merge vowel [y] with vowel [u] in terms of F3 also. BL and LL did not mark the contrast /ø – œ/ in terms of F3 (Fig. 1f). Duration was not affected by English duration patterns for neither BL nor LL.

In conclusion, subjects deviated more in English than in French. The degree of divergence varied depending on formant and language background. In English, LL deviated significantly from TV for F1. F1 values for ML and LL deviate from the TV when it comes to mid [ɛ ɔ] and low [æ ɑ] vowels. This contradicts the results from the study (Babel, 2012) which states that low vowels are more likely to be imitated than high vowels; F1 especially. LL deviation in English could be explained by socio-dialectal differences amongst speakers as stated in the study by Horton et al. (2011). BL diverged significantly in terms of F2; /ɪ, u, ɛ, æ/ were more fronted than TV. English vowel contrast /ɑ - ɔ/ and French vowel contrast /e – ɛ/ were difficult for most participants. In French, no significant differences were found. BL and language LL relied heavily on F1 for /ø - œ/ distinction. LL relied on F2 for /u – y/ distinction. The acoustic cues prioritized in English and in French by both BL and LL suggest a difference in production strategies. An articulatory study would thus prove useful. Graded levels of interaction between phonological systems for both BL and signs of transfer for the LL (Flege, 1995) present. Preliminary results from this study support Paradis' *Interactional Dual Systems Model* (Paradis, 2001). Furthermore, this study does not equate phonetic convergence to intrinsic phonological competence as some vowel contrasts were problematic for LL and BL even where divergence was not significant. In English, divergence was significant for both BL and LL but vowel contrasts were maintained in terms of formants and duration. Thus, knowing where and how variability occurs is a definitive factor where phonological competence is concerned.

References

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. Journal of Phonetics, 40(1), 177-189.

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.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance, 2(1), 125.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of phonetics*, *34*(4), 500-515.

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1-1.

Paradis, J. (2001). Do bilingual two-year-olds have separate phonological systems?. *International journal of bilingualism*, *5*(1), 19-38.

Regan, V., & Bayley, R. (2004). Introduction: The acquisition of sociolinguistic competence. *Journal of sociolinguistics*, *8*(3), 323-338.

Speech temporal control modulated by prosodic factors and sense of agency

Jinyu LI, Laboratoire de Phonétique et Phonologie (CNRS/Sorbonne-Nouvelle)
jinyu.li@sorbonne-nouvelle.fr

Flexibility of speech motor control in the temporal dimension, observed in the durations of speech gestures, is especially shown by the studies on time-delayed auditory feedback (DAF), in which speakers hear their speech with a certain delay and consequently respond to this temporal mismatch by decreasing their speech rate (i.e., lengthening syllables) (e.g., Yates, 1963; Kalveram and Jäncke, 1989). However, given the complexity of the speech motor control system, we may expect that the temporal flexibility of speech production is modulated by various factors, including prosodic factors (e.g., the syllable position in the prosodic structure) and psycholinguist factors (e.g., the sense of agency during speech production, who is a determinant factor of controlling our own speech - Welniarz et al., 2013).

To bring evidence in favor of these hypotheses, we conducted an experiment based on real-time perturbations of auditory feedback, in which 30 French speakers heard their speech with a certain delay (0, 60 or 120 ms), and/or with a shift of the F0 (0, 1 or 2 semitones). More precisely, we tested if with increased delay in the auditory feedback, accented syllabic nuclei were lengthened more than non-accented nuclei. Moreover, we tested if the constant F0 shift in the auditory feedback could alter the speakers' sense of agency during speech production, and if this effect could modulate the speaker's responses to DAF.

The results show that speakers' response to DAF depend on the syllabic status in the prosodic hierarchy. DAF lengthens more accented syllables and thus increases the saliency of accented syllables, leading to a reorganization of speech rhythm, as demonstrated by a strengthening of the coordination of syllabic and supra-syllabic amplitude modulations (see Figure1). The results also show that the constant F0 shift may affect the speakers' sense of agency, reducing thus the effects of DAF. However, this reduction effect interacts with the effect of the prosodic structure.

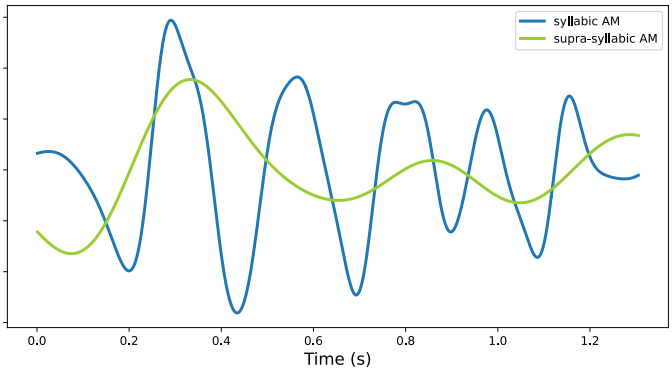


Figure 1: Example of the amplitude modulation of the sentence /ʒaklin ʒɛʁ lə ʒuʁ/

References

Kalveram, K. T., & Jäncke, L. (1989). Vowel duration and voice onset time for stressed and nonstressed syllables in stutterers under delayed auditory feedback condition. *Folia Phoniatrica*, 41(1), 30–42.

Welniarz, Q., Worbe, Y., & Gallea, C. (2021). The forward model: A unifying theory for the role of the cerebellum in motor control and sense of agency. *Frontiers in Systems Neuroscience*, 15.

Yates, A. J. (1963). Delayed auditory feedback. *Psychological Bulletin*, 60(3), 213.

Origin and consequences of linguistic stereotypes: a case study on linguistic perception of children and learners of Italian as second language

Camilla Masullo, Universitat Rovira i Virgili
camilla.masullo@urv.cat

Linguistic stereotypes are defined as the act of judging people according to their linguistic output (Lippi-Green, 1997 *inter alia*).

This project aims to investigate the concept of linguistic stereotype considering both its development in children and its effect on foreign language learning. To achieve these aims, we conducted two different studies: the first one on a child sample (Study 1), the second one on learners of Italian as second language (Study 2).

Study 1 is part of a major research project, which is focused on linguistic and cultural consequences of the migration involving the area of Biella, in Piedmont (north of Italy), and which also included the creation of CoLIMBi corpus (Corpus of Language, Identity and Migration in Biella). The purpose of Study 1 is to inquire when linguistic stereotypes develop and whether sex, age and linguistic background play a role in children's linguistic perception. To answer these questions, we collect data from 79 school pupils (44F, 35M) from 6 to 9 years old. Linguistic attitudes towards six regional varieties of Italian were elicited through the matched guise technique. In experiment one, children were asked to listen to six Italian accents (i.e. Piedmontese, Lombard, Venetian, Neapolitan, Sicilian and Sardinian) and to answer some questions about socioeconomic status of the speaker: responses were elicited by using emoticons corresponding to different Likert-scale values.

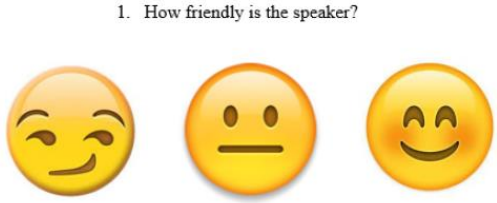


Figure 1: Emoji Likert-scale for Friendliness

In experiment two, children were asked to choose from which of the six Italian voices they would have preferred to receive a present with the purpose of indirectly eliciting their preference towards accents. Both quantitative and qualitative data analyses were led through IBM SPSS 20. The combination of different variables (child's sex, age, linguistic and migration background) showed specific patterns of children's linguistic attitudes. First, factor age played a crucial role with younger children being more affected by familiarity effect and choosing accents from their own linguistic background, and older children showing linguistic attitudes similar to the ones of adults and recognizing linguistic varieties' social prestige. Older pupils were also better at detecting prosodic and phonological differences between auditory stimuli. Additionally, with respect to sex, female students revealed to develop linguistic attitudes earlier than boys, in line with other investigations (Lambert et al., 1966; Rosenthal, 1977; Preston and Bayley, 1996 *inter alia*). Finally, children's linguistic background influenced accents' recognition since immigrant children with different spoken languages/dialects tended to easily recognize accents' variation, in line with Floccia and colleagues (2009).

Study 2 ranks among language perception and language acquisition with the aim of verifying whether foreign learners of Italian L2 show linguistic attitudes towards Italian accents and what is the role of language perception in language learning. The importance of linguistic attitudes in the framework of language acquisition has been highlighted especially for English as L2, where perception of learners' accents was found to play a crucial role (Scales et al., 2006; Kim, 2012; Kalra and Thanavisuth, 2018 *inter alia*).

The sample included 18 Spanish learners of Italian, 26 learners of Italian from other linguistic backgrounds and 6 Italian respondents as control group. Verbal guise technique

was employed: eight Italian accents for which both female and male speakers between the age of 24 and 30 were recorded were used as auditory stimuli. Respondents were asked to fulfill a questionnaire where judgements about comprehensibility of the voice and supposed social status of the speaker, besides information about respondents' experience with Italian L2 were gathered. Data from the questionnaires were analyzed both through quantitative and qualitative analysis with IBM SPSS 20.
Results showed the importance of factors such as L2 exposure and motivation on comprehension. Although specific linguistic stereotypes were not detected, a positive correlation between comprehensibility of the accent and its pleasantness was found. Results clearly show how Italian accents are perceived by foreign learners and how accents' perception has an impact on acquisition of Italian L2.

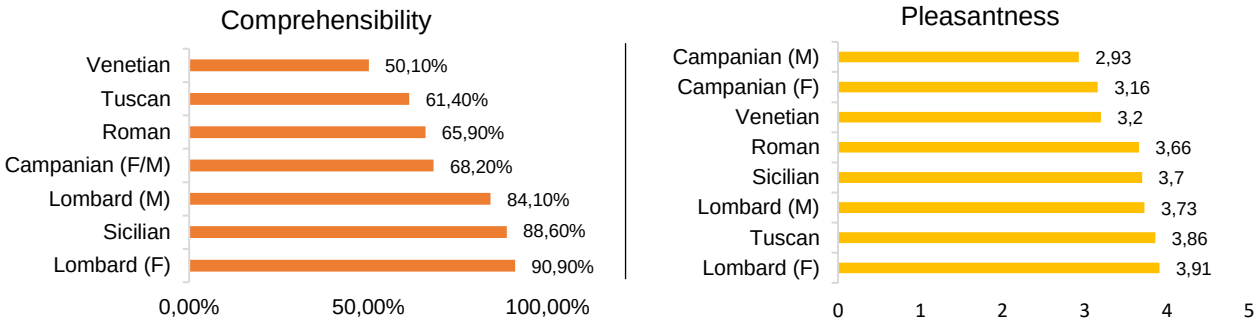


Figure 2: Judgements of comprehensibility and pleasantness of Italian accents. Comprehensibility is calculated by considering the percentages of correct answers to comprehension questions for each accent. Pleasantness is calculated on 1-5 Likert scale judgements (1 minimum, 5 maximum).

Taken together, results of Study 1 and 2 allowed us to build a thorough overview of the concept of linguistic stereotype, from its development in children to its consequences in human activities such as language learning, with potential insights on education system and foreign languages' teaching.

References

Floccia, C., Butler, J., Girard, F., and Goslin, J. (2009). Categorization of regional and foreign accent in 5- to 7-year-old British children. *International journal of behavioral development*. *33*, 366–375.

Kalra, R., & Thanavisuth, C. (2018). Do you Like My English? Thai Students' Attitudes towards Five Different Asian Accents. *Arab World English Journal*, *9*(4), 281–294.

Kim, J. (2012) English accents and L2 learner's identity. *International Journal of English and Education*, *1*(2), 127–152.

Lambert, W.E., Frankle, H., and Tucker, G.R. (1966). Judgement personality through a speech: a French-Canadian example. *Journal of Communication*, *16*, 305-321.

Lippi-Green, R. (1997). *English with an accent: Language, ideology and discrimination in the United States*. London: Routledge.

Preston, D. R., & Bayley, R. (1996). Second Language Acquisition and Linguistic Variation (Studies in Bilingualism (SiBil)). John Benjamins Publishing Company.

Rosenthal, M. S. (1977). *The magic boxes: Children and Black English*. The ERIC Clearinghouse on Languages and Linguistics/Center for Applied Linguistics: Arlington.

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language Learners' Perceptions of Accent. *TESOL Quarterly*, *40*(4), 715.

On-body radar-based silent speech interface

João Vítor Possamai de Menezes, Institute of Acoustics and Speech Communication, TU Dresden
joao_vitor.possamai_de_menezes@tu-dresden.de

Speech-based interaction between humans and machines is a reality in everyday life, mainly because of its advantages over screen- and keyboard-based interaction, e.g., portability, speed and comfort. Software such as Amazon's Alexa, Apple's Siri and Google Assistant apply deep neural networks trained with immense data sets to perform real-time large vocabulary speech recognition. These applications depend on the acoustical speech signal to function and are, therefore, severely impaired when this specific signal is partially or completely unavailable. Some of those situations are when people are physiologically unable to speak (e.g. laryngectomy patients), when confidentiality is needed (e.g. in public spaces where conversations can be overheard by others) or when the acoustic noise of the environment masks audible speech (e.g. in cockpits of aircrafts). In this context, silent speech interfaces (SSIs) are a solution that enables speech communication to take place without the acoustic signal, as they rely on speech related bio-signals such as muscle and neural activity, as well as articulatory movements.

To allow a spread use of SSIs in real life situations, these systems must be simultaneously stable (make reproducible measurements under the same circumstances), portable (as small as possible), convenient (easy to use) and non-invasive (without any sensors under the skin or inside the mouth). Many of the SSIs developed so far do not fulfill all of these requirements (Gonzalez-Lopez et al, 2020), but an on-body radar-based SSI under development in our research group has the potential to do so.

Our system works by establishing a radar with two or three antennas (one emitting and the others receiving) placed on the face of the speaker. We emit a broadband signal (1-6 GHz) through the skin into the upper vocal tract, which is received after being reflected ("filtered") by the articulators (mostly the tongue) and the upper vocal tract walls. Since phonemes have specific articulation, i.e. vocal tract geometry, the received signal varies consistently with the different phonemes.

Our baseline system works with two antennas fixed by a headset, one on each cheek of the speaker, as shown in Figure 1A. Studies are being carried out to determine where and whether to place a second receiving antenna. Our recording setup is shown in Figure 1B.

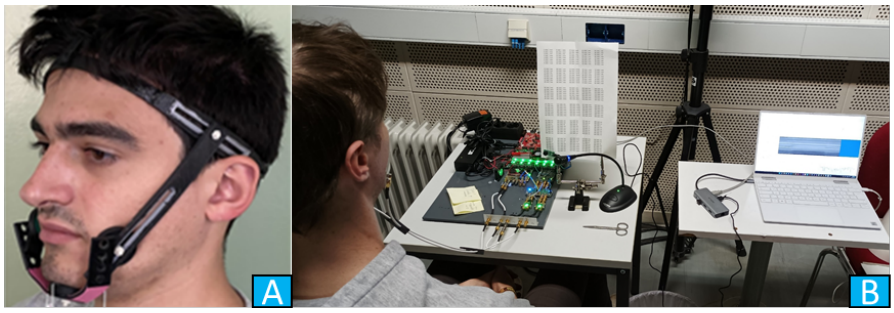


Figure 1: A) Baseline method for recording, with two antennas placed on the cheeks of the speaker. A custom headset stabilizes them (Digehsara et al, submitted). B) Recording setup. Here the radar hardware (with green LEDs) sends and receives the signals through the antennas on the speaker's face. It then processes the data and sends it to a computer, for further analysis.

The results obtained so far in classification experiments with labeled data sets recorded with our SSI are promising. The accuracies obtained in a phoneme recognition experiment were 93% and 85% for two different speakers (Birkholz et al, 2018), and the accuracies obtained in word recognition experiments were 99.17% considering data from only one session and 88.87% considering test data from unknown sessions (Wagner et al, 2022).

The potential of this technology is being discovered, which is exciting, given the good results presented so far. The next steps we plan to take are the following:

- Perform speech synthesis based on radar data (radar-to-speech);
- Perform continuous speech recognition based on radar data (radar-to-text);
- Test different signal types for the radar. What we used so far is a continuous-wave stepped frequency radar, but we are also developing an impulse radar, which should be sampled faster and easier to miniaturize;
- Investigate antenna types and positions, in order to prove whether a third antenna aids the SSI's performance.

References

Birkholz, P., Stone, S., Wolf, K. and Plettemeier, D. (2018). *Non-invasive silent phoneme recognition using microwave signals*. IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26(12), pp.2404-2411. doi: 10.1109/TASLP.2018.2865609

Gonzalez-Lopez, J. A., Gomez-Alanis, A., Martín Doñas, J. M., Pérez-Córdoba, J. L. and Gomez, A. M. (2020). *Silent Speech Interfaces for Speech Restoration: A Review*. IEEE Access, vol. 8, pp. 177995-178021. doi: 10.1109/ACCESS.2020.

Wagner, C., Schaffer, P., Amini Digehsara, P., Bärhold, M., Plettemeier, D. and Birkholz, P. (2022). *Silent speech command word recognition using stepped frequency continuous wave radar*. Scientific Reports, vol. 12, article number 4192. doi: https://doi.org/10.1038/s41598-022-07842-9

Amini Digehsara, P., Possamai de Menezes, J. V., Wagner, C., Bärhold, M., Schaffer, P., Plettemeier, D. and Birkholz, P. (submitted). *A user-friendly headset for radar-based silent speech recognition*.

Trade-offs between performance and effort during listening in adverse conditions

Alex Mepham, University of York
am2050@york.ac.uk

Speech perception in noise has been shown to be more effortful than speech perception in quiet. In addition, perceiving speech in a noisy background is usually harder when the background is intelligible to the listener, an indication of *linguistic interference*. For example, it is harder to disambiguate a talker from speech played forward (i.e., time-forward speech), than speech played backward (i.e., time-reversed speech). Pupil dilation can be used as a measure of listening effort, with tasks requiring greater cognitive effort evoking greater magnitudes of pupil dilation. However, most experiments using pupillometry as a measure of listening effort aggregate responses across many trials, which may miss granular changes over the course of sustained exposure to background noise. Typically, pupil dilation decreases over the course of an experiment as listeners learn to segregate target from masker talkers. In this study, we were interested in whether reductions in pupil dilation are similar across intelligible and unintelligible competing talkers, and whether changes in speech recognition performance are reflected in changes in listening effort. To test whether listeners can learn to control linguistic interference, and to explore listening effort deployed in speech recognition in noise, we tracked their transcription ability and mean pupil dilation against time-forward and time-reversed English maskers over 50 trials. We established the intensity level leading to 50% correct for each participant in the time-forward and time-reversed conditions using an adaptive procedure, so that all participants began the speech recognition task at the same performance level. The average signal-to-noise (SNR) ratio needed to achieve 50% correct was approximately 2 dB higher in the time-forward than time-reversed condition, confirming linguistic interference from time-forward speech. In the recognition task, performance improved faster in the time-forward than time-reversed maskers (Figure 1). In contrast, mean pupil dilation did not show a difference between the time-forward and time-reversed conditions; performance decreased at similar rates in both conditions (Figure 2). We interpret the faster improvement in performance in the time-forward condition as evidence that a meaningful masker, although intrinsically more likely to generate linguistic interference (hence the higher initial SNR), is more amenable to segregation over time than a meaningless masker. In other words, although a time-forward masker can lead to heightened linguistic interference initially, it becomes easier to isolate and inhibit as participants gain more exposure to the maskers over time. The lack of difference between masker conditions in the pupillometric data suggests that time-related improvement in the segregation of a meaningful masker does not necessarily lead to a corresponding decrease in effort, pointing to important trade-offs between performance and effort during listening in adverse conditions.

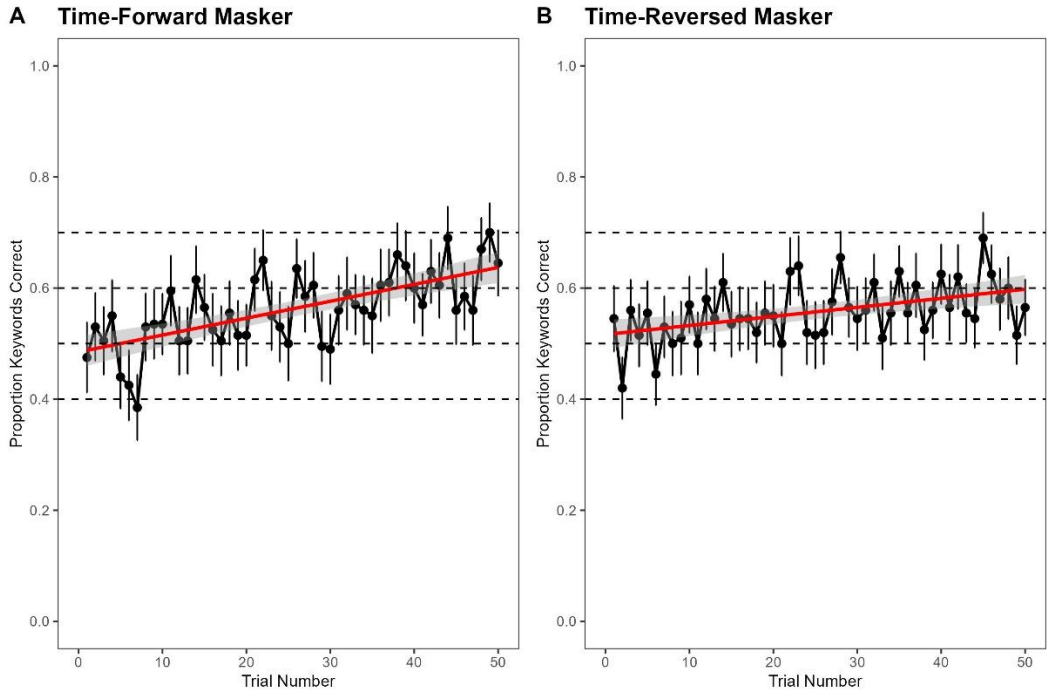


Figure 1. Mean proportion of keywords correctly reported for each trial in each Masker condition. Each dot is the mean proportion of keywords correctly reported for each trial. The error bars are the standard error of the mean (SE) for each trial, with the red lines representing the linear trend across the Masker condition. The dashed horizontal line indicates performance at 40%, 50%, 60% and 70% correct.

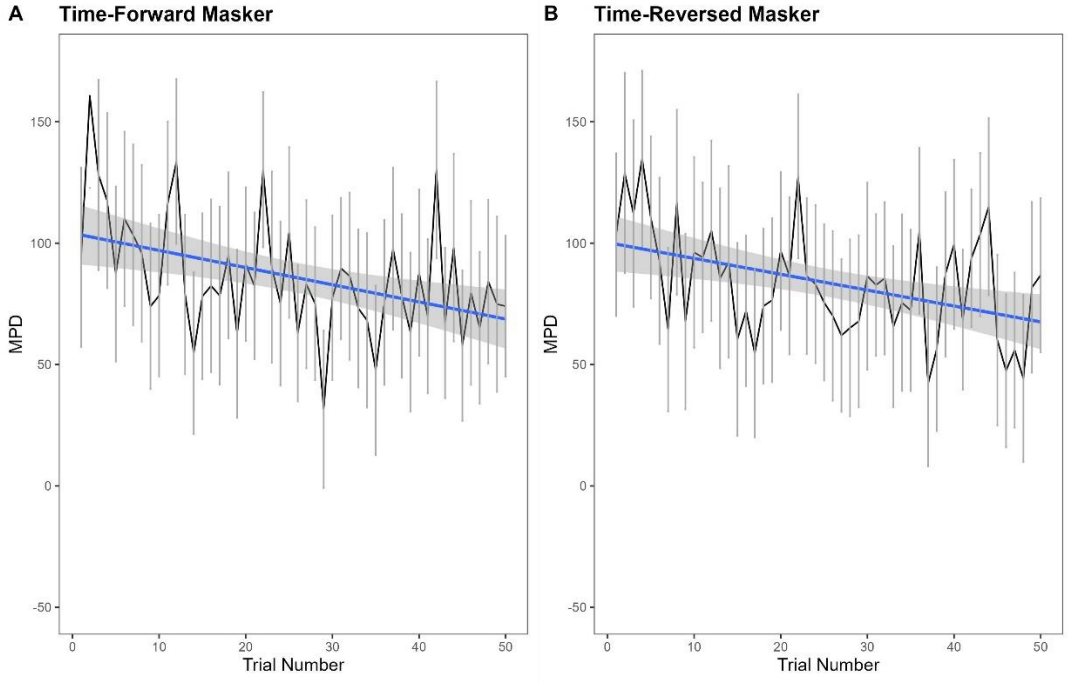


Figure 2. Mean pupil dilation (MPD) for each trial in each experimental Block. For each trial, the MPD value is the mean MPD across participants for each trial. The error bars are the standard error of the mean (SE) for each trial, with the blue lines representing the linear trend across the Masker condition.

A lightweight physics-based vocal tract acoustic model using the finite-difference time-domain method

Debasish Ray Mohapatra, University of British Columbia
Victor Zappi, Northeastern university
Sidney Fels, University of British Columbia
debasishray@ece.ubc.ca

The human voice is a complex but unique physiological process. It involves the neuromuscular control of articulators to form an intricate upper vocal tract geometry resulting in different speech sounds. The classic articulatory speech synthesizer represents the vocal tract geometry as juxtaposed cylindrical tube segments with different cross-sections, and then it employs an acoustic wave solver that numerically approximates wave propagation inside these tube segments to synthesize speech utterances. Although this approach closely follows the speech production mechanism, it requires a high computational overhead. With the emerging medical imaging technologies and the explosion of computing capabilities, the modern 3D articulatory models (Arnela & Guasch, 2013) can precisely represent vocal tract geometries and compute their acoustic characteristics. However, these models are still far from synthesizing speech in real-time. Alternatively, the wave solver can bind wave propagation to a single dimension (1D articulatory models) for faster acoustic simulation by assuming planar wave propagation inside a straight cylindrical tube. However, the radial symmetry of the 1D acoustic tube (van den Doel, 2008) only describes the emergence of fundamental modes while eliminating transverse modes. Therefore, the oversimplified 1D vocal tract models can only precisely approximate the lower formants (up to 5kHz) of the speech spectrum.

We propose a novel vocal tract acoustic model (i.e., 2.5D FDTD vocal tract (Mohapatra et al., 2019)) that extends the existing 2D model (Zappi et al., 2016) while having acoustic characteristics comparable to 3D models. The model defines the 2D mid-sagittal vocal tract contours inside a computational domain to estimate wave propagation. Additionally, it lumps the effect of nonplanar waves across the mid-sagittal plane by mapping tube depth (i.e., tube height across the sagittal plane derived from vocal tract area functions) to the 2.5D acoustic wave solver. The inclusion of tube depth reduces the linear acoustic wave equations to the 2D Webster's horn equation for the computation of wave propagation inside the vocal tract geometry. Unlike the existing 2D vocal tract models, this approach eliminates the previous requirement of nonlinear scaling of vocal tract area functions. We have adapted the local reactive boundary approach for the mid-sagittal contour to account for the vocal tract boundary losses. We have also incorporated losses due to wall vibration by considering vocal tract boundaries as elastic walls. This technique offers an excellent balance between the computational cost and acoustic precision while promising better geometrical flexibility for the vocal tract modelling. The discretized wave solver equations for the 2.5D FDTD vocal tract model has been described as follows,

$$p^{n+1} = \frac{\overline{D}p^n - \rho c^2 \Delta t(\widetilde{\nabla}.\boldsymbol{V}^n)}{\overline{D}}$$

$$v^{n+1} = \frac{\beta v^n - \beta^2 \Delta t\, \widetilde{\nabla} p^{n+1}/\rho \; + \Delta t\,(1-\beta) v_b}{\beta + \Delta t(1-\beta)}$$

with:

$$\boldsymbol{V} = (D_x v_x, D_y v_y)$$

where p is the sound pressure and v is the particle velocity. The ρ and c represent the air density and sound speed respectively. The term $\overline{D}$, D_x and D_y are the components of the depth map of the 2.5D space. A detailed derivation of depth map from vocal tract area function has been provided here (Mohapatra et al., 2019).

This study demonstrates the synthesis of cardinal vowels using the 2.5D FDTD vocal tract model and characterizes their acoustic features through transfer function analysis. We used a highly precise 3D FEM vocal tract model as the baseline model to compare the acoustic formants. For both the FDTD and FEM models, we used Story's 1D area function dataset (Story, 2008) to generate vocal tract models.

Table 1: Percentage positional error of the first eight formants in 2.5D FDTD model, computed for vowels */a/, /i/* and */u /* with respect to the 3D FEM model.

	/a/	***/i/***	***/u/***
f_1	-1.16%	1.48%	2.70%
f_2	1.12%	0.69%	4.01%
f_3	0.22%	0.74%	0.13%
f_4	0.29%	0.72%	-0.66%
f_5	-0.34%	0.48%	-1.02%
f_6	0.17%	-0.15%	1.79%
f_7	1.06%	-0.18%	1.25%
f_8	-0.41%	0.69%	-1.04%

Table 1 shows the percentage positional error of acoustic formants between the 2.5D FDTD and 3D FEM vocal tract models. Overall, results show good agreement with errors that stay below 4%. Vowel /a/ and /i/ produce the best matches, while the first two formants of vowel /u/ displayed a more significant shift. As the next step, we are currently working on the inclusion of the free-radiation effect using perfectly matched layers (PMLs). We also plan to explore the possibilities of modelling various simplified geometries using the 2.5D vocal tract model with different degrees of complexity, such as straight tubes with elliptical cross-sections and bent tubes with circular and elliptical cross-sections, etc.

References

Arnela, M., & Guasch, O. (2013). *Finite element computation of elliptical vocal tract impedances using the two-microphone transfer function method.* The Journal of the Acoustical Society of America, 133(6), 4197-4209.

van den Doel, K., & Ascher, U. M. (2008). *Real-time numerical solution of Webster's equation on a nonuniform grid.* IEEE transactions on audio, speech, and language processing, 16(6), 1163-1172.

Zappi, V., Vasuvedan, A., Allen, A., Raghuvanshi, N., & Fels, S. (2016). *Towards real-time two-dimensional wave propagation for articulatory speech synthesis.* In Proceedings of Meetings on Acoustics 171ASA (Vol. 26, No. 1, p. 045005). Acoustical Society of America.

Mohapatra, D. R., Zappi, V., & Fels, S. (2019). *An extended two-dimensional vocal tract model for fast acoustic simulation of single-axis symmetric three-dimensional tubes.* Proceedings of Interspeech.

Story, B. H. (2008). *Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002.* The Journal of the Acoustical Society of America, 123(1), 327-335.

Do filler particles facilitate the recollection of lists?

Beeke Muhlack, *Language Science and Technology (Saarland University)*
muhlack@lst.uni-saarland.de

Filler particles, such as *uh* and *um* in English, are frequently used in spontaneous speech. There is some evidence that these filler particles may have benefits for the listener such as improvements on the recollection of words (Corley et al. 2007; Collard et al. 2008). Fraundorf and Watson (2011) found that filler particles that occur in discourse improve the recollection of important events. The participants were better at recalling short passages of *Alice in Wonderland* when the story included filler particles in contrast to those who heard the story without filler particles. However, in a previous web-based study (Muhlack et al. 2021) we were not able to replicate this finding using English and German data. This suggests that the beneficial effect of filler particles on recollection may be rather small. The reduced attention of subjects in a web-based study may have further reduced the recollection effect.

To investigate this effect further, we designed an experiment using lists of six different categories (i.e. body parts, clothing, fruit, musical instruments, vegetable, zoo animals). The lists included high and low frequency items, that were determined in a pretest beforehand. Each list consists of twelve items from the same category, six high frequency words and six low frequency words. Before two items (1 high and 1 low frequency item) of each list the filler particle *um* occurs, which is manually spliced into the recording using Praat. Participants listen to each list and are asked to recall the items they heard by writing them down in an answer box on the screen. If filler particles indeed benefit recall, we expect that participants are better at recalling the items that were preceded by a filler particle. The experiment was created using LabVanced, 73 native speakers of English are recruited via Prolific. The results do not show a beneficial effect of the filler particles on the recall of lists but rather a recency effect is observed in addition to an effect of the subjects' memory capacity.

References

Collard, P., Corley, M., MacGregor, L.J. & Donaldson, D.I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 34, no.3, pp. 696-702.

Corley, M., MacGregor, L.J. & Donaldson, D.I. (2007). It's the way you, er, say it: Hesitations in speech affect language comprehension, Cognition, vvol. 105, pp. 658-668.

Fraundorf, S.H. & Watson, D.G. (2011). The disfluent discourse: Effects of filled pauses on recall, Journal of Memory and Language, vol. 65, no.2, pp.161-175.

Muhlack, B., Elmers, M., Drenhaus, H., Trouvain, J., van Os, M., Werner, R., Ryzhova, M. & Möbius, B. (2021). Revisiting recall effects of filler particles in German and English. Interspeech 2021. Brno, Czechia.

Comprehension, production and processing of Maltese noun inflections: A modeling approach using LDL

Jessica Nieder, Yu-Ying Chuang, Ruben van de Vijver & Harald Baayen, Heinrich-Heine Universität Düsseldorf, Eberhard-Karls-Universität Tübingen
nieder@phil.hhu.de

The Discriminative Lexicon (DL) (Baayen et al., 2019) is a recent model of the mental lexicon that makes use of Linear Discriminative Learning (LDL) and its computations, which is a mathematical formalization of Word and Paradigm Morphology and its assumption that words and their inflectional paradigms are the basic units of morphological analysis (Blevins, 2016). The DL model sets up mappings between phonological form and meaning, without requiring prior analysis of forms into sequences of stems and morphemes or exponents. At the same time, the model assumes that the meanings of words are constructed from the meaning of the content word and the inflectional meanings that are realized in a word's form. The DL model has been used to model the comprehension and production of complex words, and it has also been used to generate predictions for lexical processing.

In this study, we use LDL to model the production and comprehension of Maltese nouns, and to obtain a better understanding of the results of a cross-modal priming study of Maltese nouns reported in Nieder et al. (2021b).

Maltese, a Semitic language spoken in Malta, shows a bewildering amount of concatenative sound plurals, e.g. *annimal-annimali* 'animals', and non-concatenative broken plurals, e.g *kelb-klieb* 'dogs'. Previous investigations have shown that speakers in wug tasks can be quite unsure as to how to form plurals from singulars (Nieder et al., 2021a). Thus, the noun morphology of Maltese appears to have limited productivity.

The LDL models that we constructed for the comprehension and production of Maltese nouns were able to learn mappings between form and meaning, and meaning and form, with high accuracy when evaluated on the training data: 99.9% for comprehension and 96.3% for production. When tested on held-out data, the model showed a lowered accuracy 73% for comprehension and 69% for production. This modeling result dovetails well with the above mentioned finding that speakers are highly reluctant to create plurals for unseen nouns.

We also investigated whether the LDL models might help us understand better previous results obtained with a cross-modal priming study, in which auditory plural primes preceded written singular targets. For this study, we made use of semantic vectors (aka word embeddings) generated with Fasttext (Joulin et al., 2016).

We compared a baseline Generalized Additive Model (Wood, 2017) with the predictors word frequency and word length (following Nieder et al. (2021b)) with an alternative model using measures from the LDL analysis. The LDL model provided an improved fit.

We conclude that the DL framework provides a promising tool for predicting the accuracy with which Maltese plural nouns can be understood and produced on the one hand, and the processing costs of these words as gauged with a primed cross-modal lexical decision task.

References

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Blevins, J. P. (2016). The minimal sign. In A. Hippisley & G. Stump (Eds.), *The Cambridge Handbook of Morphology* (pp. 50–69). Cambridge University Press.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv*. https://arxiv.org/abs/1612.03651

Nieder, J., van de Vijver, R., & Mitterer, H. (2021a). Knowledge of Maltese singular–plural mappings. *Morphology*, 31, 147–170.

Nieder, J., van de Vijver, R., & Mitterer, H. (2021b). Priming Maltese plurals: Representation of sound and broken plurals in the mental lexicon. *The Mental Lexicon*, 16(1), 69–97.
Wood, S. N. (2017). *Generalized Additive Models*. Chapman & Hall/CRC.

Does CASE trump determiners? Considering blocking effects in heritage Turkishes in Germany and the U.S.

Onur Özsoy, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
oezsoy@leibniz-zas.de

Languages apply a wide range of strategies to mark definiteness and specificity. In many IE languages (e.g., German, English, Greek), (in)definiteness is expressed via determiners, articles and demonstratives. Turkish lacks a definite article and employs accusative case to mark definiteness and specificity. For NPs which contain the accusative marker but not the indefinite article bir 'one', a definite interpretation is assumed (1) (von Heusinger & Kornfilt, 2005). Some also argue that Turkish is an article-less language (Bošković & Şener, 2014). However, it may use demonstratives to signal a definite interpretation.

(1) (bu) kitab-ı oku-du-m
this book-ACC read-PRF.PST-1SG
'I read the / this book.'

In contact situations, marking of definiteness may be affected by language transfer dynamics (Polinsky, 2006). Definiteness in heritage Turkishes is interesting to investigate as it is an under-researched field in the study of heritage languages, especially in the context of two different majority languages, namely German and English. Thus, we explore Turkish heritage speakers' strategies for the expression of definiteness and whether language contact leads to the emergence of new linguistic patterns in this domain. Further, the design of our study allows to capture some effects of multimodality in language production, as participants were narrating into different devices (smartphone vs face-to-face) and the interlocutors were dressed differently (police-like, formal vs home-like, informal) depending on the setting.

Data from two age groups of heritage and monolingual Turkish speakers were elicited via a narration task in Germany, the U.S. and Turkey. In total, data from 186 speakers was collected. The stimulus was a video of a mild fictional car accident which participants narrated in two different modes (oral and written) and two communicative situations (to a close friend, informal, and to the police, formal) (Wiese, 2020). We find that heritage speakers generalize the accusative-marking strategy (β = -0.16, z(19750) = -4.18, p < .001) and pattern monolingual-like regarding the use of alternative strategies of marking definiteness in Turkish, i.e., a blocking effect where CASE comes before demonstratives is possible.

Our findings are partly explained by language contact effects but also point to internal dynamics in the development of the accusative in heritage Turkishes. Emerging patterns of definiteness-marking in heritage languages call for analyses which have implications for ongoing theoretical discussions, e.g., the status of determiners and specificity in Turkish (Hedberg et al., 2009). Thus, we will conclude by revisiting some of the discussions about definiteness marking in (heritage) Turkishes (Erguvanlı-Taylan & Zimmer, 1994; Felser & Arslan, 2019; Kamali, 2015; Kupisch et al., 2017).

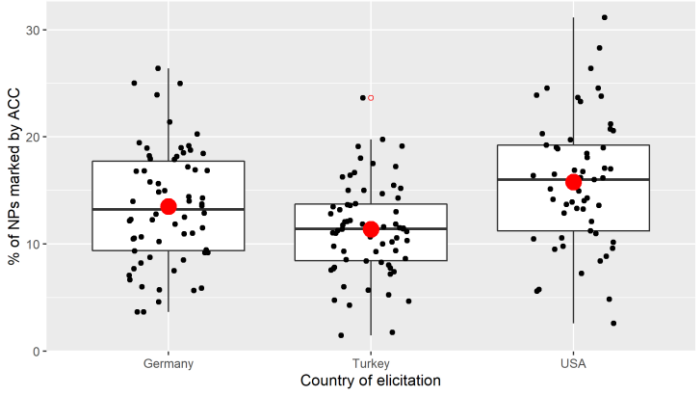


Figure 1: Boxplots representing the % of ACC-marked NPs by participant per group.

References

Bošković, Z., & Şener, S. (2014). The Turkish NP. *Crosslinguistic studies on noun phrase structure and reference* (pp. 102–140). Brill.

Erguvanlı-Taylan, E., & Zimmer, K. (1994). Case marking in Turkish indefinite object constructions. *Annual Meeting of the Berkeley Linguistics Society*, 20 (1), 547–552.

Felser, C., & Arslan, S. (2019). Inappropriate choice of definites in Turkish heritage speakers of German. *Heritage Language Journal*, 16 (1), 22–43.

Hedberg, N., Görgülü, E., & Mameni, M. (2009). On definiteness and specificity in Turkish and Persian. *Proceedings of the 2009 Annual Meeting of the Canadian Linguistic Association*.

Kamali, B. (2015). Caseless direct objects in Turkish revisited. *ZAS Papers in Linguistics*, 58, 107–123.

Kupisch, T., Belikova, A., Özçelik, Ö., Stangen, I., & White, L. (2017). Restrictions on definiteness in the grammars of German-Turkish heritage speakers. *Linguistic Approaches to Bilingualism*, 7 (1), 1–32.

Polinsky, M. (2006). Incomplete acquisition: American Russian. Journal of Slavic linguistics, 191–262.

von Heusinger, K., & Kornfilt, J. (2005). The case of the direct object in Turkish: Semantics, syntax and morphology. *Turkic languages*, 9, 3–44.

Wiese, H. (2020). Language situations: A method for capturing variation within speakers' repertoires. Methods in Dialectology, 16.

Vowel Perception in Congenital Amusia

Jasmin Pfeifer, Heinrich-Heine-Universität & Silke Hamann, Universität von Amsterdam
pfeifer@phil.hhu.de

Congenital amusia is a disorder that negatively influences pitch and rhythm perception (e.g. Peretz et al. 2002) and is not caused by a hearing deficiency or brain damage. While congenital amusia had long been reported to affect only the musical domain (Peretz et al. 2002, Ayotte et al. 2002), several studies have shown that amusics also have impaired perception of intonation (e.g. Patel et al. 2008) and linguistic tones (e.g. Tillmann et al. 2011).

In the present study we tested whether congenital amusia also has an influence on linguistically relevant cues other than pitch by investigating the discrimination of German front high vowels /ɛ/, /ɛ:/, /e/ and /e:/ (where /ɛ:/ and /e/ are of low frequency). We assessed amusics' behavioural and electrophysiological responses, more specifically the Mismatch negativity (MMN), which is a component evoked by unconscious change detection in the auditory signal (e.g. Näätänen 2001).

We tested 11 congenital amusics diagnosed with the MBEA (Peretz et al. 2003) and 11 matched controls. All participants were right handed, had normal hearing and had German as their native language. Our stimuli were isolated synthetic vowels varying in either duration or spectral properties. We used synthesized vowels to ensure very tightly controlled stimuli (Iverson 2012), while at the same time, we tried to keep them as naturally sounding as possible by adding a falling-rising pitch contour and amplitude, and eight additional formants. We decided to use mid vowels to avoid periphery effects (Polka & Bohn 2003), and to utilize vowels that are close to each other in their height and front-back dimension in the vowel space, but that differ in quality and/or quantity.

In the behavioural study we employed an ABX task. The stimuli were presented with an inter-stimulus interval (ISI) of either 0.2 s or 1.2 s.

The stimuli in the EEG study were presented in a multi-deviant oddball paradigm in four blocks. In each block, one of the four high front vowels was the standard and occurred 85% of the time, while the other three served as deviants, each occurring 5% of the time. This resulted in 16 event-related potentials (ERPs) per participant: 4 standards and 12 deviants. The inter-stimulus interval was varied randomly between 400 ms and 600 ms to avoid entrainment effects.

For the behavioral data, we calculated a linear mixed model (lmer in R) with subject as random effect, and group (amusics vs. controls), ISI (0.2 s vs. 1.2 s) and cue (duration vs. formant frequency) as fixed factors. We found main effects of group ($t(20)=2.28$, $p=0.033$), ISI ($t(1028)=7.69$, $p<0.001$) and cue ($t(1028)=8.24$, $p<0.001$). As expected, amusics performed worse than controls. Furthermore, short ISI resulted overall in worse performance, and a difference in duration was overall harder to detect than a difference in formant frequency.

For the MMN data, we used a linear mixed model as well. We found significant main effects for group ($t(23.7)=-2.43$, $p=0.023$): amusics ($M=-2.67$) had a smaller MMN than controls ($M=-3.35$) visible in Figure 1 (right panel). In addition we found a main effect for cue ($t(2351.8)=-6.14$, $p<0.001$) and a significant interaction between group and cue ($t(2351.8)= 3.85$, $p<0.001$): durational differences were harder to detect, especially for amusics.

Our study shows that congenital amusia does not only affect the perception of pitch in music and language but also the perception of vowel contrasts, therefore having more far-reaching consequences for speech perception than previously assumed. Not only was the behaviour of amusics shown to be affected, we also found differences in the MMN, reflecting differences in early auditory change detection.

References

Peretz, I., Ayotte, J., Zatorre, R., Mehler, J., Ahad, P., Penhune, V., & Jutras, B. (2002). Congenital Amusia: A Disorder of Fine-Grained Pitch Discrimination. Neuron, 33, 185–191.

Ayotte, J., Peretz, I., & Hyde, K. 2002. Congenital amusia - A group study of adults afflicted with a music-specific disorder. Brain, 125, 238–251.

Patel, A., Wong, M., Foxton, J., Lochy, A., & Peretz, I. 2008. Speech intonation perception deficits in musical tone deafness (congenital amusia). Music Perception, 25, 357–368.

Tillmann, B., Burnham, D., Nguyen, S., Grimault, N., Gosselin, N., & Peretz, I. 2011. Congenital amusia (or tone-deafness) interferes with pitch processing in tone languages. Frontiers in Psychology, 2. doi: 10.3389/fpsyg.2011.00120

Näätänen, R. 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent MMNm. Psychophysiology 38, 1–21.

Peretz, I., Champod, S., & Hyde, K. 2003. Varieties of Musical Disorders: The Montreal Battery of Evaluation of Amusia. Annals of the New York Academy of Sciences, 999, 58–75.

Iverson, P. (2012). Measuring phonetic perception in adults. In A. C. Cohn, C. Fougeron & M. Huffman (Eds.), The Oxford Handbook of Laboratory Phonology (pp. 572-580). Oxford: Oxford University Press.

Polka, L., & Bohn, O.-S. (2003). Asymmetries in vowel perception. Speech Communication, 41(1), 221-231

Warm + fuzzy: Perceptual semantics can be activated even during surface lexical processing

Olesia Platonova, Potsdam Embodied Cognition Group, University of Potsdam; Laboratory of Linguistic Anthropology, Tomsk State University
pla.olesia@gmail.com
Alex Miklashevsky, Potsdam Embodied Cognition Group, University of Potsdam,
armanster31@gmail.com

Understanding language relies on sensorimotor areas of the brain (Hauk, 2004; Pecher, 2003). Previous studies demonstrated that sequential processing of words related to particular perceptual modalities in context (e.g., “roses can be *red*”, “ice can be *cold*” or “blender can be *loud*”) leads to activation of corresponding sensorimotor brain structures (i.e., *vision*-, *touch*- or *audition*-related).

Neurocognitive studies (Bernabeu, 2017; Hald, 2011) revealed activation of modality information preceding the peak of semantic recognition. However, the depth of the semantic processing and the broadness of the context necessary for the activation of the sensorimotor brain areas remains not fully understood (Cayol, 2020).

In our study we investigated modality activation during shallow semantic processing without a context.

Method. 68 native Russian participants were assessing 180 pairs of simultaneously presented stimuli (‘word + word’, for example, ‘warm + fuzzy’ or ‘word + pseudoword’, for example, ‘yellow + kemily’). Participants were asked to perform a lexical decision task for both stimuli at once and give a response by pressing a key. A total of 360 stimuli were presented in random order, 180 counterbalanced pairs per participant, half contained at least one pseudoword. The linguistic material of the experiment was selected from the psycholinguistic adjective database (Kolbeneva, 2010). Stimuli set consisted of 72 adjectives from three semantic modalities (visual, tactile, or auditory - for example, white, warm, or quiet, respectively). Stimulus pairs were formed by enumerating options for combining modalities, for example, "modality_1 - modality_1" and "modality_2 - modality_1".

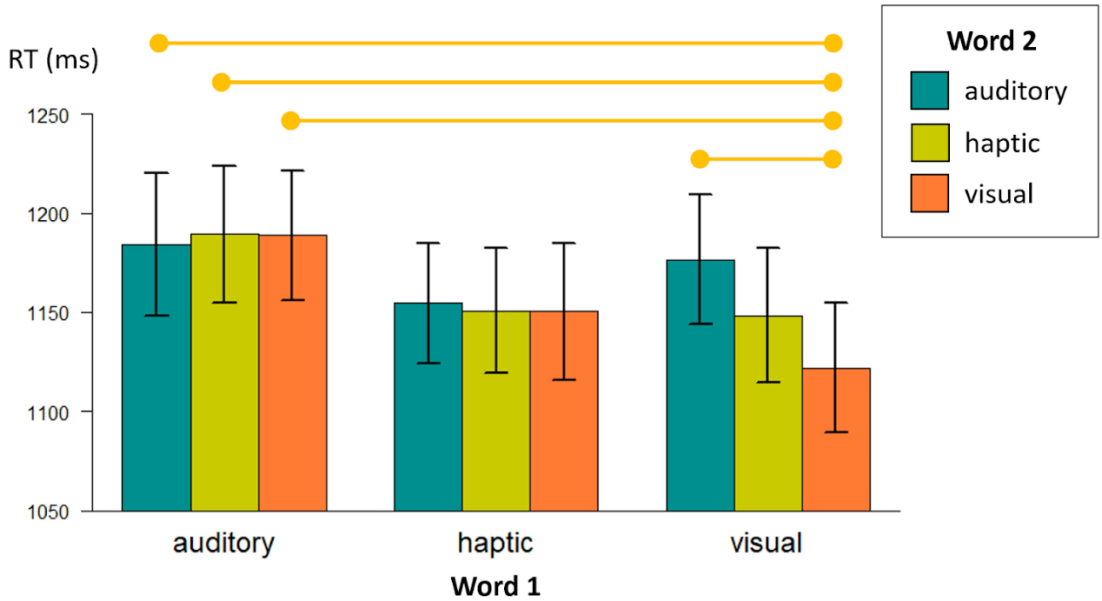


Figure 1: Mean reaction times for nine conditions (combinations of semantic modalities: Modality 1 X Modality 2; see main text for details). Whiskers represent standard errors. Orange horizontal lines represent significant differences between conditions.

Results. When the first word represented the visual modality, reaction times differed according to the modality of the following word. Combination of modalities ‘visual + visual’ turned out to be significantly faster than ‘visual + audial’ ($p = .022$) or combination of the audial

modality with any other modality (p-values < .004). No significant differences were found between combinations of visual and tactile modalities.

Our study demonstrates that even surface lexical processing without a context is able to activate perceptual semantics.

References

Bernabeu, P., Willems, R.M., Louwerse, M.M. (2017). Modality Switch Effects Emerge Early and Increase throughout Conceptual Processing: Evidence from ERPs. *CogSci 2017 Proceedings*, 1629-1634.

Cayol, Z., Nazir, T.A. (2020). Why Language Processing Recruits Modality Specific Brain Regions: It Is Not About Understanding Words, but About Modeling Situations. *Journal of Cognition*, *3,* 1–23.

Hald, L.A., Marshall, J.A., Janssen, D.P., Garnham, A. (2011). Switching Modalities in A Sentence Verification Task: ERP Evidence for Embodied Language Processing. *Frontiers in Psychology*, *2*.

Hauk, O., Johnsrude, I., Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron, 41.* 301–306.

Kolbeneva, M.G., & Aleksandrov, Y.I. (2010). *Sense organs, emotions and adjectives of the Russian language: Lingvo-psychological dictionary*. Moscow: Languages of Slavic cultures.

Pecher, D., Zeelenberg, R., Barsalou, L.W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science Research Article,14*. 119–124.

Time-varying spectral characteristics of Danish stop releases

Rasmus Puggaard-Rode, Leiden University Centre for Linguistics
r.p.hansen@hum.leidenuniv.nl

The aspirated alveolar stop /t/ in Modern Standard Danish is affricated quite dramatically, but does not pattern phonologically as an affricate. This has led to a variety of different descriptions and transcription strategies. /p k/ are usually taken to be 'regular' aspirated stops. No systematic phonetic studies have ever been made of these stops' release characteristics, and it is not clear if the affrication in /t/ is simply mentioned more often because sibilant frication is most salient. There are no obvious acoustic phonetic heuristics for determining whether a sound is an affricated stop or an actual affricate, but with the right tools, examining the spectrum can go some way towards answering whether affrication is environmentally determined, speaker-specific, etc.

There are a number of methods in use for analyzing frication on the basis of the spectrum. Calculating spectral moments – i.e. treating the complex spectrum as a probability mass function – can be used to summarize the spectrum (Forrest et al., 1988). Particularly the spectral mean (or center of gravity) is a popular measurement in this regard. The mid-frequency spectral peak can pinpoint relatively minor differences in place of articulation quite precisely (Koenig et al., 2013), as can the four highest cepstral coefficients of a discrete cosine transform (Bunnell et al., 2004). These measurements are usually taken at static or normalized points in time ('magic moments'; see Mücke et al., 2014). There are advantages and disadvantages to all these approaches, but perhaps the main advantage of all of them is that they reduce the complex, multi-dimensional information in the spectrum into a manageable number of discrete values, which can then serve as dependent variables in statistical models.

In this poster, I propose function-on-scalar regression (FOSR; Bauer et al., 2018) as a method to model spectra and their variation in time holistically and without the need to reduce dimensionality. In functional data analysis, variables do not need to be discrete values, but can instead be (smoothed) curves, such as spectra. An FOSR model is similar to a linear mixed-effects model or a generalized additive mixed model, except with a functional response instead of a discrete continuous response. The output of FOSR models give clear and easily interpretable overviews of the influence of various factors on the functional response. Examples are given in Figures 1–2. In Figure 1, we see the influence of time on the energy distribution in the spectrum; the structure is similar to spectrograms, with normalized time along the x-axis, frequency along the y-axis, and grey-scale shading indicating energy. Figure 1 shows significant energy at frequencies above 5 kHz early on in /t/ releases, and significant energy below 5 kHz towards the end of the release. This suggests that the stop releases are initially affricated, but that this affrication is gradually dominated by aspiration. Figure 2 shows how this pattern differs for male and female speakers.

For this study, I used the spontaneous monologues in the DanPASS corpus (Grønnum, 2009). For each aspirated stop (n = 2,334), multitaper spectra were extracted from 20 normalized time steps. Separate FOSR models were fitted for each phoneme /p t k/, and the influence of each dependent variable subsequently visualized. The results show that /t/ is indeed invariably affricated, although the spectrum is very dynamic throughout /t/ releases, with affrication gradually being replaced by aspiration. The acoustic characteristics of /p/ releases show a lot of inter-speaker variation, but also coarticulatory context effects, primarily during the first half of the release. Coarticulatory context effects greatly influence the spectra of /k/ releases throughout. These findings can all be explained with reference to general acoustic and articulatory principles, and with reference to previous studies of Danish stop articulation, in particular muscular and glottal activity (Fischer-Jørgensen and Hirose, 1974; Hutters, 1985). I believe that using FOSR models to analyze time-varying spectra could potentially be fruitful in the analysis of many problems in acoustic phonetics beyond this one.

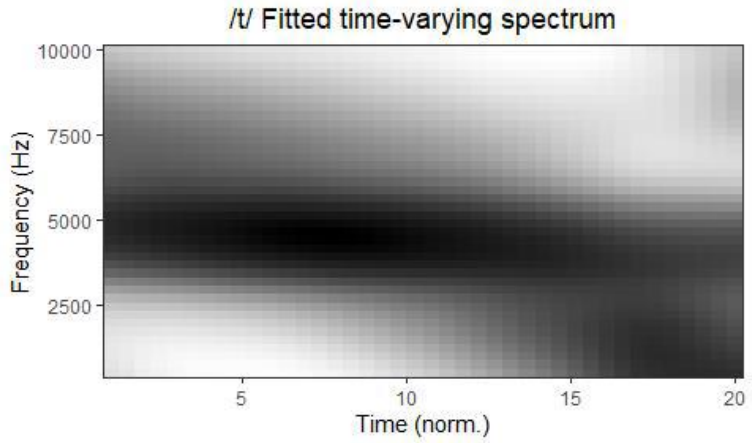


Figure 1: Fitted time-varying spectrum of /t/ releases (main effect of time).

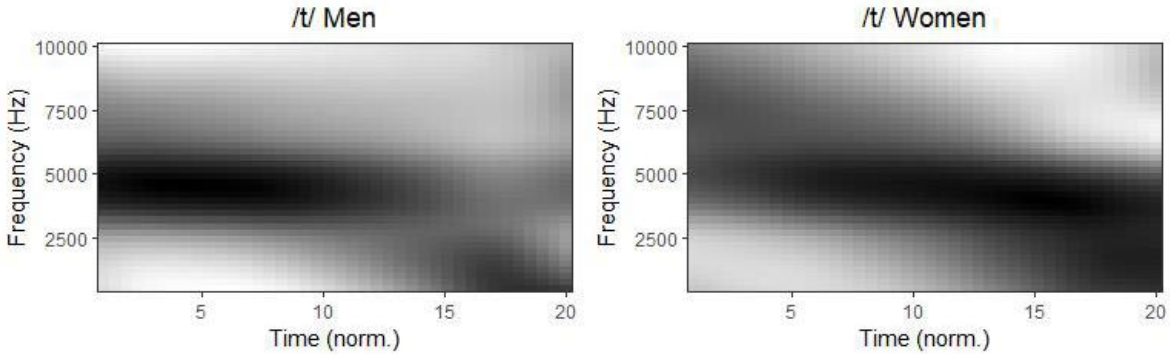


Figure 1: Fitted time-varying spectrum of /t/ releases for each direction of the sex variable.

References

Bauer, A., F. Scheipl, H. Küchenhoff & A.-A. Gabriel (2018). An introduction to semiparametric function-on-scalar regression. *Statstical Modelling, 18*, 346–364.

Bunnell, H.T., J. Polikoff & J. McNicholas. (2004). Spectral moment vs. bark cepstral analysis of children's word-initial voiceless stops. *International Conference on Spoken Language Processing, 8*.

Fischer-Jørgensen, E. & H. Hirose. (1974). A preliminary electromyographic study of labial and laryngeal muscles in Danish stop consonant production. *Status Report on Speech Research, 39/40*, 231–253.

Forrest, K., G. Weismer, P. Milenkovic & R.N. Dougall. (1988). Statistical analysis of word-initial voiceless obstruents. Preliminary data. *Journal of the Acoustical Society of America*, *84*, 115–123.

Grønnum, N. (2009). A Danish phonetically annotated spontaneous speech corpus (DanPASS). *Speech Communication, 51,* 594–603.

Hutters, B. (1985). Vocal fold adjustments in aspirated and unaspirated stops in Danish. *Phonetica, 42,* 1–24.

Koening, L.L., C.H. Shadle, J.L. Preston & C.R. Mooshammer. (2013). Toward improved spectral measures of /s/. Results from adolescents. *Journal of Speech, Language, and Hearing Research, 56,* 1175–1189.

Mücke, D., M. Grice & T. Cho. (2014). More than a magic moment. Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics, 44*, 1–7.

TRAJECTORY FORMATION IN SPEECH PRODUCTION: DOES OPTIMALITY MATTER?

Ny Tsiky Rakotomalala, Gipsa-lab
ny-tsiky.rakotomalala@grenoble-inp.fr,

Optimal control theory is one of the leading theories in motor control, but it has rarely been used to study the control of articulatory models of speech production, probably due to the complexity of the biomechanics of the orofacial system. Nevertheless, the method is potentially powerful since it can predict whole trajectories from a set of goals and a simple cost (like neuromuscular effort) to be minimized.

We aim to confront this theory with the hypothesis extensively tested by one of us, that biomechanics have a decisive influence on the shape of trajectories (Perrier P., Payan Y., Zandipour M. & Perkell J., 2003). To this end, we compare vowel production from two different models: the GEPPETO model (Patri J.-F., Diard J. & Perrier P., 2015; Payan, Y., & Perrier, P., 1997), in which actual trajectories result from the specification of sensory targets and the biomechanical characteristics of the speech production system, and an optimal control model in which trajectories are determined by the selection of optimal motor command patterns, on top of biomechanical constraints. A direct comparison of the results is presented in the acoustic domain, in the kinematic domain and in the motor domain.

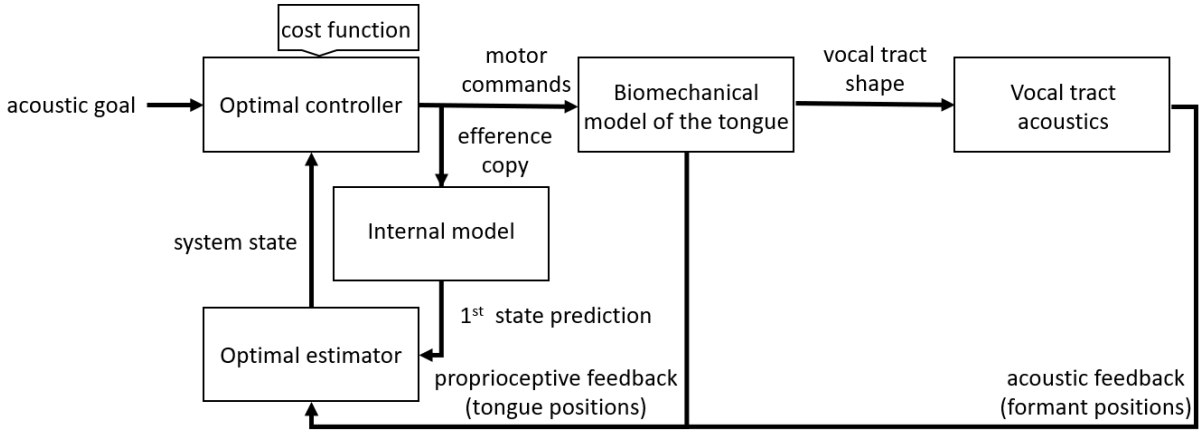


Figure 1: Architecture of the optimal control model.

References

Perrier P., Payan Y., Zandipour M. & Perkell J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants : A modeling study. The Journal of the Acoustical Society of America, 114(3), 1582–1599.

Patri J.-F., Diard J. & Perrier P. (2015). Optimal speech motor control and token-to-token variability : a bayesian modeling approach. Biological Cybernetics, 109(6), 611–626.

Payan, Y., & Perrier, P. (1997). Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech communication*, *22*(2-3), 185-205.

Articulatory flexibility following oral cancer treatment: Outline of a longitudinal study

Thomas B. Tienkamp, University of Groningen, The Netherlands & University Medical Centre Groningen, The Netherlands
Rob J.J.H. van Son, University of Amsterdam, The Netherlands & Netherlands Cancer Institute, The Netherlands
Sebastiaan A.H.J. de Visscher, University Medical Centre Groningen, The Netherlands
Max J.H. Witjes, University Medical Centre Groningen, The Netherlands
Martijn Wieling, University of Groningen, The Netherlands & Haskins Laboratories, USA

Email address: t.b.tienkamp@rug.nl

In 2018, over 400,000 people were diagnosed with cancer of the oropharynx, oral cavity, and lips worldwide (Bray et al., 2018). Tumor extension and treatment often leads to speech distortion (De Bruin et al., 2009). In order to compensate for these anatomical changes, patients need to come up with new articulatory configurations to produce similar acoustic outputs (i.e., form motor equivalence strategies). While we know from studies that the human language system is flexible in adapting to perturbations in the short term, little is known about long-term adaptation. Moreover, although some acoustic studies suggest that patients exhibit compensatory behaviour, it remains largely unknown what this behaviour looks like articulatorily and how articulatory strategies change over time. In turn, this missing knowledge contributes to the near absence of standardised speech therapy guidelines.

The objective of this project is to longitudinally investigate the coordination and development of speech articulation of patients who underwent surgical treatment for T1-T2 tumors in the oral cavity to assess: (1) to what extent oral cancer patients are able to form motor equivalence strategies; (2) how strategies develop over time; and (3) whether adaptation to auditory or tactile perturbation can predict the success of the motor equivalence strategies in the long term.

This study is a longitudinal prospective study which will start mid 2022. Evaluation points include: pre-treatment, and 6-, 12-, and 18-months post-treatment. Before treatment, reliance on auditory and tactile feedback will be tested using auditory and tactile feedback perturbation experiments. Additionally, speech will be collected using electromagnetic articulography (EMA). After surgery, only speech with EMA will be collected to decrease participant burden.

We hypothesise that compared to control speakers, oral cancer patients will show deviant articulatory trajectories and acoustics, but that they will form motor equivalence strategies in order to maximise intelligibility. Moreover, while adaptation will co-depend on the size and site of the tumor, we also hypothesise that individual differences in adaptation to auditory and tactile perturbation may predict the rate of post-treatment adaptation. If this is indeed the case, it could be explored whether feedback perturbation could be employed as a therapeutic tool.

References

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394–424. https://doi.org/10.3322/caac.21492

De Bruijn, M. J., Ten Bosch, L., Kuik, D. J., Quené, H., Langendijk, J. A., Leemans, C. R., & Verdonck-de Leeuw, I. M. (2009). Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. Folia Phoniatrica et Logopaedica, 61(3), 180-187. DOI: 10.1159/000219953.

Cognitive semantic and cognitive semantic associations among nouns of entities across English and Chinese

Yufang Wang[1], Jurriaan Witteman[1], Niels O. Schiller[1]
1. Leiden University Centre for Linguistics (LUCL)
y.f.wang@hum.leidenuniv.nl; N.O.Schiller@hum.leidenuniv.nl; j.witteman@hum.leidenuniv.nl

Background:

How do the brains of human beings with different native languages code information about entities and their relationships? As we know, nouns for entities across languages seldom have a one-to-one relationship, for example, both "祖母(grandmother)" and "外祖母(maternal grandmother)" in Chinese are usually translated into "grandmother" in English. While the associations between "king-queen" and "father-mother" are congruent across languages.

Among the existing semantics theories, context-sensitive grammar is one of the most widely accepted and used in language-related areas (Carlos, 1999). This theory argues that the meaning of a word is largely dependent on its surrounding words, e.g. "love" in "I love music" is decided by "I" and "music". However, context-sensitive grammar cannot be used to explain the relatedness of the same word in different contexts. Word2vec used this theory as its linguistic basis and fixed this problem to by giving the same initial value for the same word to obtain the distributed representation of all words according to their contexts (Mikolov, 2013). In this case, the vector transcoded by the same word is a combination of all the meanings it has. This way of transcoding was confirmed to be high similarities to human semantic networks in the light of network/graph theory (Kajic, 2018). The human semantic network was built on the task requiring participants to list nouns from the same semantic category. Semantic category refers to entities of nouns with the same function, group and/or scales. For instance, "arm" and "leg" are both body parts and are viewed as the same semantic category. This concept is also widely used in other psycholinguistic tasks, e.g. picture-word interference (PWI), a paradigm used to investigate the recruited linguistic features of the target and distractor words during language production, and received robust biological evidence(Bürki, 2020). Congruency between the word2vec and human semantic network as well as biological evidence during PWI tasks agree with the distributed representation for object-noun encoding of human brains and language (Rogers, 2021) as well as tools of language for the cognitive processing of human beings (Sirbu, 2015).

On the other hand, few studies were done on the encoding of objects and their associations across languages. Very few one-to-one relationships of cross-linguistic nouns for entities or different contexts considering the grammatical or syntactic differences would make the distributed representation for the nouns across languages different. But the similar events human beings trying to describe would yet make the encoding of the associations among nouns for entities be of significant similarity, which is also in line with the cross-language biological evidence for the same criterion of nouns for entities from the same category [8]. Together, we propose the existence of cross-language semantic representation but a shared semantic relationship across languages.

In this study, we will try to test whether there is a significant difference in the distributed representation of nouns for entities and significant correlations of associations among the nouns for entities between English and Chinese in the light of linguistic and biological. Distributed representation data/vectors, trained on large-scale data with closer distribution of the real situation, of nouns for entities, verified at a biological level, in English and Chinese will be used for this study. Then cosine angle/distance among the obtained vectors will be used to measure semantic associations among nouns for entities.

Method:

We reused the word list of the study by Bürki et al., which summarized the words with robust biological evidence with the same cross-language criterion of semantic category based on a meta-analysis of previous experiments using picture-word interference (Bürki, 2020). The vectors we used in this study of these words were trained on Common Crawl and Wikipedia using fastText by Grave et al. (Grave, 2018). Got vectors of these words, cosine distance, a

widely accepted way for language associations, was recruited to measure the semantic relations. Obtaining the semantic associations, paired two-sample t-test and a correlation test will be used to test whether there is a significant difference for vectors for entities and significant correlations of associations among the nouns for entities between English and Chinese.

Results:

(1). Significant difference ($t = -273.8$, $p < 0.05$) is shown between cosine angles among recruited English nouns and that of Chinese.

(2). No significant difference but a correlation with a medium effect size (Pearson's $r = 0.34$, $p = 0$) cosine angles of vectors for recruited nouns are shown between English and Chinese.

Conclusion:

This significant difference of semantics and congruency of semantic relations among nouns for entities between English and Chinese might suggest that (1) variations of noun encoding/distributed semantic representation across languages; (2) a shared basis for semantic associations among nouns for entities across languages; (3) Meanwhile, it might provide some biological evidence for the building of word2vec model as well.

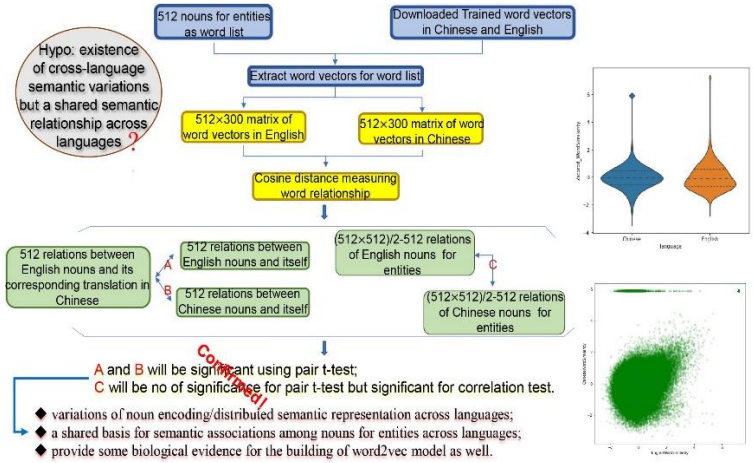


Figure 1: This is the procedure and the results of the study.

References

Carlos Martín Vide, ed. (1999). Issues in Mathematical Linguistics: Workshop on Mathematical Linguistics, State College, Pa., April 1998. John Benjamins Publishing. pp. 186–187. ISBN 90-272-1556-1

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Kajic, I., & Eliasmith, C. (2018). Evaluating the psychological plausibility of word2vec and GloVe distributional semantic models. Technical Report CTN-TR-20180824-012. Centre for Theoretical Neuroscience, University of Waterloo. Waterloo, ON, Canada. doi: 10.13140/RG. 2.2. 25289.60004.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. Journal of Memory and Language, 114, 104125.

Rogers, T. T., Cox, C. R., Lu, Q., Shimotake, A., Kikuch, T., Kunieda, T., ... & Ralph, M. A. L. (2021). Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. eLife, 10.

Sirbu, A. (2015). The significance of language as a tool of communication. Scientific Bulletin" Mircea cel Batran" Naval Academy, 18(2), 405.; Tylén, K., Weed, E., Wallentin, M., Roepstorff, A., & Frith, C. D. (2010). Language as a tool for interacting minds. Mind & Language, 25(1), 3-29.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Breath noise formants in speech production

Raphael Werner, Saarland University
rwerner@lst.uni-saarland.de

This project deals with the acoustics of speech breathing. In particular, it is concerned with formants present in inhalation noises in speech pauses. In a data set, originally described in Rochet-Capellan & Fuchs 2013, where 34 female German native speakers retold short fables, audible breath noises were annotated into being either right before speech (*inh-ini*) or sandwiched between stretches of speech (*inh*). Some speech segments that could share similarities with inhalations were annotated as well, such as /ə/ to locate inhalations with regard to a neutral setting of the vocal tract and /ɐ, a:/ to compare inhalations with more open segments (similar to Werner et al. 2021 but adding more reference sounds). Further, we annotated /i:, u:/ for reference.

For the first part, formants were measured as averaged over the mid third, i.e. from 33 % to 67 % of their duration, to avoid potential influence of coarticulation with surrounding speech. Comparing the formants of breath noises to those of the aforementioned speech segments, inhalations appear to have a higher F1 than /ə/ and are in that respect more similar to /ɐ/ but most similar to /a:/ (see Fig. 1). However, inhalations' F2 tends to be higher than those of /ɐ, a:/. Locating them on a vowel chart, they thus appear more open than /ə/ and more fronted than /ɐ, a:/.

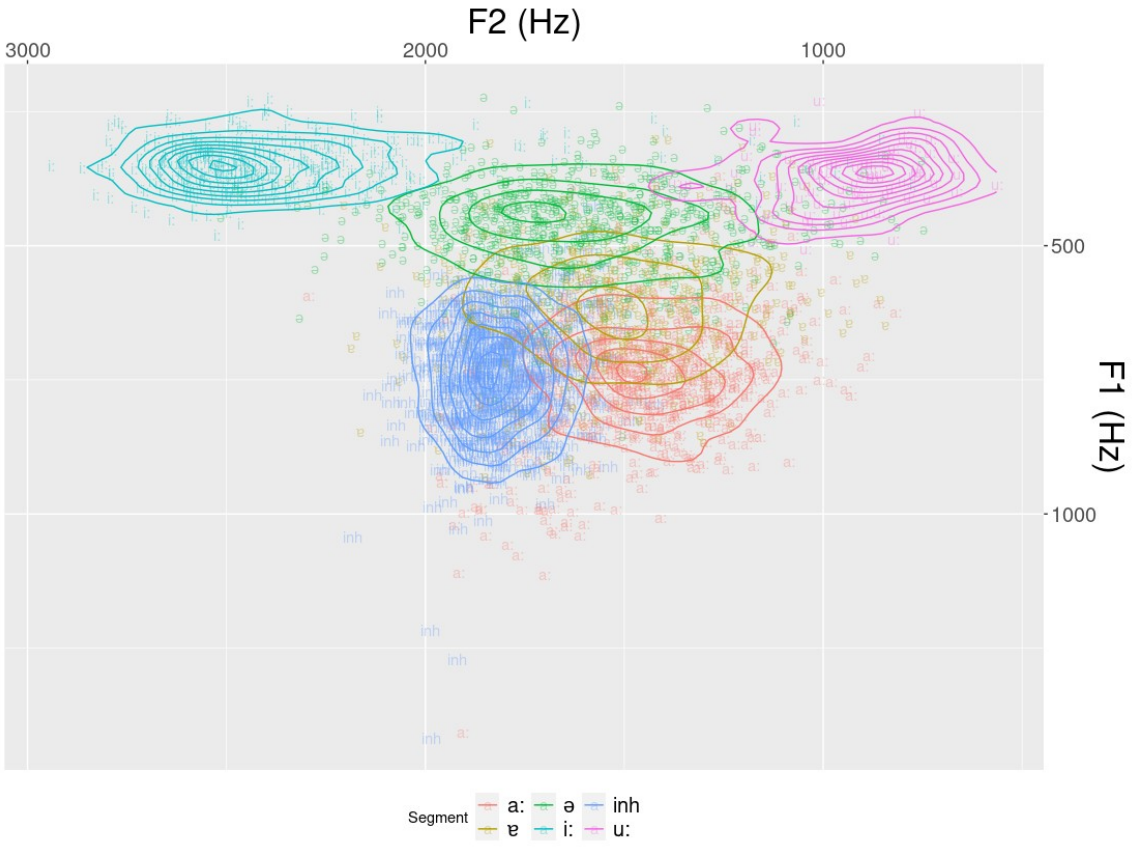


Figure 1: Vowel chart showing first and second formant as averaged over the mid third for all included sounds in Hertz, as well as their 2D-density. Inhalations (*inh*) are in blue, /ə/ in green, and /ɐ, a:/ are in red and yellow.

In the second part, I would like to present data from the same set but here formants for breath noises were not measured as averages over a part of the segment, or even midpoints

only, but values extracted at 11 points from 0 % to 100 % of the respective duration. I used Generalized Additive Mixed Models (GAMMs) to plot F1, F2, and F3 for the two speech inhalation types, namely *inh-ini* and *inh*. Analyzing the formant trajectories allows to accommodate for the dynamic nature of formants and helps to discover changes over time but it also shows potential differences between the inhalation types and if and where they have a stable region to justify averaging over, as done in the previous section. Preliminary analyses suggest that differences in shape between the two types are found mainly for F1 and partly for F2. Along with shape, F1 also seems to have the biggest differences in height between the two types. For F3, shapes are similar but there seem to be height differences. These differences might be related to *inh* tending to be shorter or potentially more nasal participation in inh-ini. Overall, the GAMMs have a relatively stable region in the middle third, thus using averages over that part for other analyses seems justifiable.

Among the things I would like to discuss are how to test the acoustic findings with articulatory studies. In particular, I am interested in how to advance from here, e.g. can real-time MRI be used or is a different type of data better suited, and how well can velum and jaw movement be examined there? Investigating jaw movement is motivated by this study and velum movement by learning more about its role and thereby nasal involvement in speech breathing (cf. Lester & Hoit 2014). Importantly, what are other ways of safeguarding against false inferences from the audio? Since formants are not as clear as they are in phonated speech and the first formant especially becomes quite weak occasionally, can bandwidth help here? Furthermore, I would like to get some feedback on GAMMs for breathing formants: are they appropriately applied, what can we infer from them? Does some of the movement at their beginnings and endings indicate coarticulation?

References

Lester, R. A. and J. D. Hoit, “Nasal and oral inspiration during natural speech breathing,” Journal of Speech, Language, and Hearing Research., vol. 57, no. 3, pp. 734–742, 2014.

Rochet-Capellan, A. and S. Fuchs, “Changes in breathing while listening to read speech: the effect of reader and speech mode,” Frontiers in psychology, vol. 4, p. 906, 2013.

Werner, R., Fuchs, S., Trouvain, J., & Möbius, B. Inhalations in Speech: Acoustic and Physiological Characteristics. Interspeech 2021, 3186–3190. 2021.

Illusions of ungrammaticality in foreign accented speech perception

Sarah Wesolek, Leibniz-Centre General Linguistics (ZAS), Berlin
Marzena Zygis, Leibniz-Centre General Linguistics (ZAS), Berlin
Piotr Gulgowski, Leibniz-Centre General Linguistics (ZAS), Berlin
wesolek@leibniz-zas.de

Communicating with foreign accented speakers can have negative consequences for the processing of linguistic information. When listeners are exposed to foreign accented speech, their general processing slows down (e.g., Braun et al., 2011; Clarke & Garret, 2004). Foreign accented speech has been experimentally linked to problems with word recognition (Bradlow & Pisoni, 1999) and a reduced anticipation of the upcoming linguistic information (Schiller et al., 2020). Furthermore, foreign accent can evoke associations and stereotypes about the speaker (Lev-Ari & Keysar, 2010). Crucially, studies indicated that a foreign accent modulates the processing of grammatical and semantic anomalies in spoken language (e.g., Hanuliková et al., 2012; Grey & van Hell, 2017). An interesting and yet underexplored phenomenon associated with the processing of foreign accented speech is the "grammatical tinnitus", a phenomenon reflecting the perception of non-existent grammatical errors in foreign accented sentences. In the present poster, behavioral results from an ERP study will be presented in which grammaticality judgments of foreign (compared to native) accented sentences were utilized to investigate if listeners perceive illusions of non-existent grammatical errors when listening to foreign accented speech.

To test the grammatical tinnitus under laboratory conditions, we conducted two mirror ERP experiments in Germany and Poland. 33 (17 female, 16 male) German monolinguals took part in the German experiment and 30 (16 female, 14 male) Polish monolinguals in the Polish experiment. All participants were asked to listen to sentences from their native language that were either foreign accented or native accented. Additionally, one third of sentences in both conditions was well formed, another third contained a phonological substitution. The remaining third contained a grammatical mistake. After listening to a given sentence, participants were asked to answer the question "Is this sentence grammatically correct?" by pressing the button corresponding to "Yes" or "No".

Our results based on a binomial logistic regression with Accent [foreign, native], Sentence type [well-formed (WF), grammatically incorrect (GRAM), phonologically incorrect (PHON)] and their interaction as well as their random structure reveal that foreign accented well formed (WF) sentences are more likely to be rated as being grammatically incorrect than native accented sentences in both German (z=3.44, p<0.01) and Polish (z=3.76, p<0.001). Additionally, in the Polish experiment foreign accented sentences that contained a phonological substitution, but no grammatical errors (PHON), were more likely to be rated as being ungrammatical than native accented sentences (z=3.76, p<0.001).

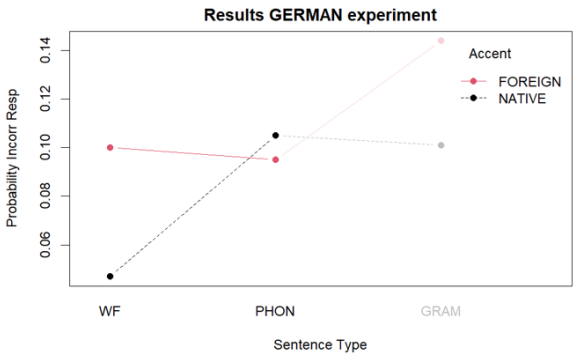


Figure 1: Probability of incorrect responses per Sentence Type and Accent Type, German experiment

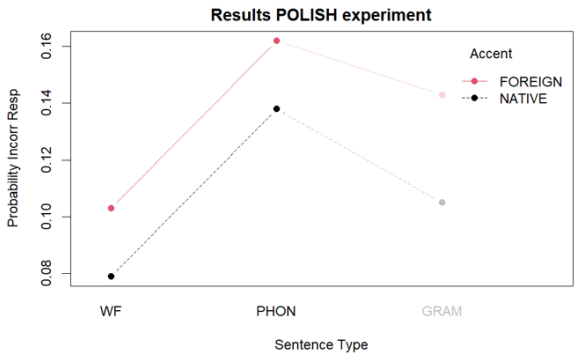


Figure 2: Probability of incorrect responses per Sentence Type and Accent Type, Polish experiment

Our results reveal that the exposure to foreign accented speech induces a perception of non-existent grammatical errors. The study will be complemented with the electrophysiological correlates of the grammatical tinnitus and listeners attitudes towards the speakers' nationality and their familiarity with the speakers' foreign accent.

References

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, 106(4 Pt 1), 2074–2085.

Braun, B., Dainora, A., & Ernestus, M. (2011). An unfamiliar intonation contour slows down online speech comprehension. *Language and Cognitive Processes*, 26, 350–375.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647–3658.

Grey, S., & van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *Journal of Neurolinguistics*, 42, 93–108.

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, 24(4).

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093–1096.

Schiller, N. O., Boutonnet, B. P.-A., De Heer Kloots, M. L. S., Meelen, M., Ruijgrok, B., & Cheng, L. L.-S. (2020). (Not so) Great Expectations: Listening to Foreign-Accented Speech Reduces the Brain's Anticipatory Processes. *Frontiers in Psychology*, 11.

Articulation of metronome-timed speech in people who stutter

Charlotte Wiltshire, Ludwig-Maximilians-University, Munich
c.wiltshire@phonetik.uni-muenchen.de

Several studies indicate that people who stutter show greater variability in speech movements than people who do not stutter, even when the speech produced is perceptually fluent (Jackson, Tiede, Beal, & Whalen, 2016; Smith, Sadagopan, Walsh, & Weber-Fox, 2010; Wiltshire, Chiew, Chesters, Healy, & Watkins, 2021). Speaking to the beat of a metronome reliably increases fluency in people who stutter, regardless of the severity of stuttering (Boutsen, Brutten, & Watts, 2000; Chesters, Watkins, & Möttönen, 2017; Frankford et al., 2021). Here, we aimed to test whether this fluency-enhancer also reduces articulatory variability. We scanned the vocal tracts of 34 people who stutter and 22 controls using MRI while participants repeated sentences with or without a metronome. Sentences contained the target words: *artillery*, *catastrophe*, *impossibility* (Ogar et al., 2006). For this analysis, we compared data from 10 people per group. Midsagittal images of the vocal tract from lips to larynx were reconstructed at 33.3 frames per second (Wiltshire et al., 2021). Any utterances containing dysfluencies or other non-speech movements were excluded. For each participant, we measured the variability of movements from the alveolar and palatal regions of the vocal tract. In previous studies, variability has been measured using two contrasting methods: (1) The coefficient of variation (CoV; Wiltshire et al., 2021), calculated by dividing the standard deviation of the summed amplitude of the raw signal, by the mean and (2) the Spatial Temporal Index (STI; Smith, Goffman, Zelaznik, Ying, & McGillem, 1995), calculated by normalizing the movement traces in time and amplitude before summing the standard deviation at a sample of 50 time points (Smith et al., 1995). Here, we calculate both measures and compare them.

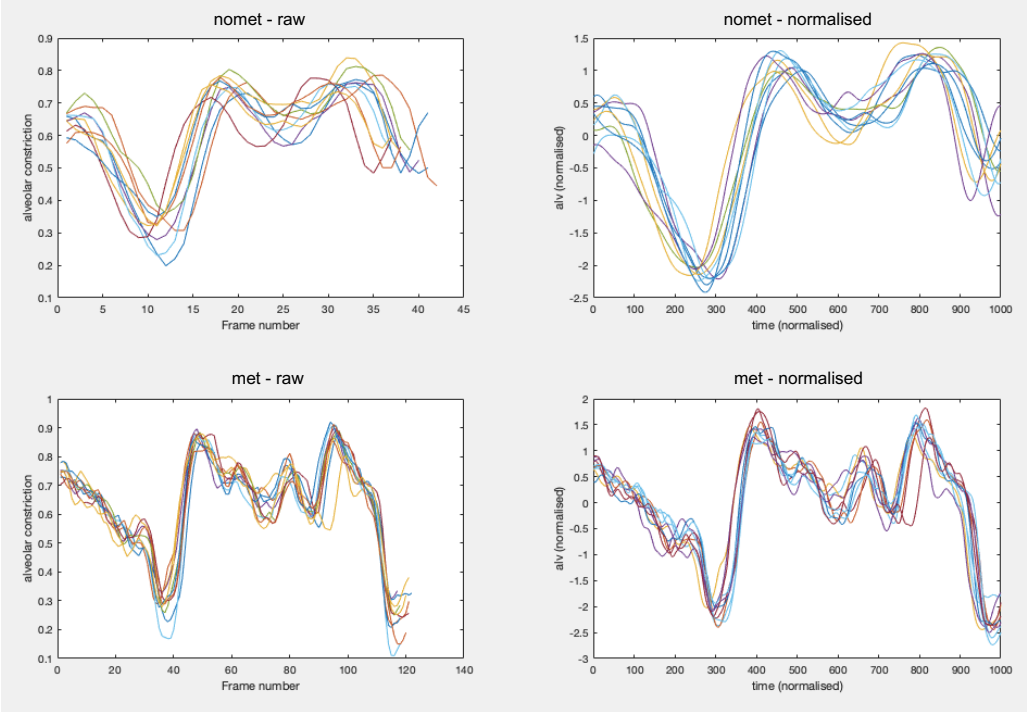


Figure 2. Example movement traces from one (stuttering) participant. Ten repetitions of "*the impossibility of*" from "*Bob knew the impossibility of the task*" are overlayed. The left panel shows raw data, used for the coefficient of variation. The right panel shows time- and amplitude-normalized data used for the Spatio-Temporal-Index. The top row shows the no-metronome condition and the bottom panel shows the metronome condition.

In line with our predictions, there was an interaction between group and condition ($f(1)=8.6$, $p=.003$) such that people who stutter had more variability than control speakers during no-metronome speech, which was then reduced to the same level as controls when speaking with the metronome.

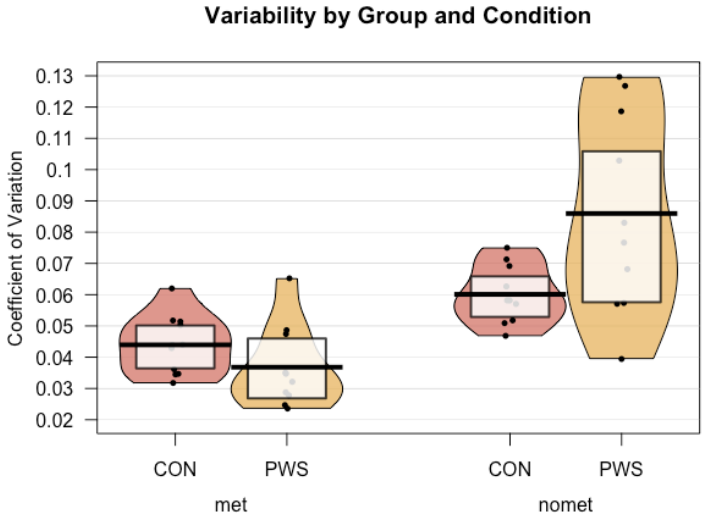


Figure 1. Mean variability (coefficient of variation) of movements in the alveolar region for control speakers and people who stutter during metronome and no-metronome speaking conditions.

In addition, there was a strong correlation between STI and CoV measures of variability ($r(128) = .66$, $p < .0001$). There was no difference in variability between alveolar and palatal regions of the vocal tract ($p = .41$). These results replicate previous findings of greater variability in the movements of people who stutter compared with controls during normal speaking (no-met condition). Furthermore, we show that the fluency enhancer, metronome timed speech, reduces variability in people who stutter to same level as control speakers.

References

Boutsen, F. R., Brutten, G. J., & Watts, C. R. (2000). Timing and Intensity Variability in the Metronomic Speech of Stuttering and Nonstuttering Speakers. *Journal of Speech, Language, and Hearing Research*, *43*(2), 513–520. https://doi.org/10.1044/jslhr.4302.513

Chesters, J., Watkins, K. E., & Möttönen, R. (2017). Investigating the feasibility of using transcranial direct current stimulation to enhance fluency in people who stutter. *Brain and Language*, *164*, 68–76. https://doi.org/10.1016/j.bandl.2016.10.003

Frankford, S. A., Heller Murray, E. S., Masapollo, M., Cai, S., Tourville, J. A., Nieto-Castañón, A., & Guenther, F. H. (2021). The Neural Circuitry Underlying the "Rhythm Effect" in Stuttering. *Journal of Speech, Language, and Hearing Research*, *64*(6S), 2325–2346. https://doi.org/10.1044/2021_JSLHR-20-00328

Jackson, E. S., Tiede, M., Beal, D., & Whalen, D. H. (2016). The impact of social–cognitive stress on speech variability, determinism, and stability in adults who do and do not stutter. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1295–1314. https://doi.org/10.1044/2016_JSLHR-S-16-0145

Ogar, J., Willock, S., Baldo, J., Wilkins, D., Ludy, C., & Dronkers, N. (2006). Clinical and anatomical correlates of apraxia of speech. *Brain and Language*, *97*, 343–350. https://doi.org/10.1016/j.bandl.2006.01.008

Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., & McGillem, C. (1995). Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research*, *104*(3), 493–501. https://doi.org/10.1007/BF00231983

Smith, A., Sadagopan, N., Walsh, B., & Weber-Fox, C. (2010). Increasing phonological complexity reveals heightened instability in inter-articulatory coordination in adults who stutter. *Journal of Fluency Disorders*, *35*(1), 1–18. https://doi.org/10.1016/j.jfludis.2009.12.001

Wiltshire, C. E. E., Chiew, M., Chesters, J., Healy, M. P., & Watkins, K. E. (2021). Speech Movement Variability in People Who Stutter: A Vocal Tract Magnetic Resonance Imaging Study. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2438–2452. https://doi.org/10.1044/2021_JSLHR-20-00507

When syntax needs prosody: How French prosodic cues help Chinese L2 learners parse syntactic information – a perception study

Lei Xi, Laboratory of Phonetics and Phonology (UMR 7018, CNRS – Sorbonne Nouvelle)
lei.xi@sorbonne-nouvelle.fr

Prosodic boundary is marked by the presence of several acoustic cues, such as pitch, final lengthening and pause. These acoustic cues are used differently in different languages. A core question in speech acquisition research is how learners exploit L2 prosodic cues to constrain syntactic ambiguity.

In this study, 40 French noun phrases belonging to two different syntactic categories (direct object or subject) were inserted within locally ambiguous sentences that differed in late / early closure (e.g., Whenever the snake {was eating the rat, the rabbit would hide. (LC) / was eating, the rat would hide. (EC)}). These sentences were produced by a French male speaker. Acoustic analyses showed that he produced reliable French prosodic cues to differentiate the ambiguous meanings. Forty intermediate (n=20, ma: 22, sd: 2.7) and advanced (n=20, ma: 24.2, sd: 2.4) Chinese L2 learners were tested whether they could correctly assign the ambiguous items (here "the rat") to their syntactic categories based on available prosodic cues. The sentences were cut after the target nouns and divided into 2 blocks: each member of a given pair appeared in a different block. In each block, we had 20 experimental stimuli (10 LC and 10 EC) plus 5 filler sentences. Participants were asked to listen to each stimulus (e.g., "Whenever the snake was eating (,) the rat") and to complete it by writing the rest. The completed sentences were coded as to whether ambiguous items were interpreted as subjects or as direct objects.

Our results showed that Chinese learners had difficulties in solving the syntactic ambiguity: they gave more LC responses (i.e., ambiguous items interpreted as direct objects) both in LC (9.85 correct responses in advanced learners vs. 9.80 in intermediate learners) and in EC (6.15 vs. 5.65 in advanced vs. intermediate learners) conditions. Two ANOVAs were conducted on the number of correct responses given in LC and EC, with proficiency level (intermediate vs. advanced) as between-participants factor and blocks as within-subject factor (block 1 vs. 2). The analyses revealed that neither proficiency level ($F(1,36)=0.17$, $p=0.68$ for LC; $F(1,36)=0.19$, $p=0.67$ for EC) nor block ($F(1, 36)=1.53$, $p=0.22$ for LC; $F(1, 36)=0.27$, $p=0.61$ for EC) had significant effect on the scores obtained.

Our results could be interpreted within the Informative Boundary Hypothesis (IBH) (Carlson et al., 2009a, b; Clifton et al., 2002) and the Late Closure Preference (Frazier, 1979). According to the IBH, the effectiveness of a prosodic boundary is determined by its size relative to relevant earlier and global prosodic boundaries in the utterance. Our stimuli were rather short (about 3-4s) and incomplete, which did not give enough prosodic information. The lack of prosodic context made learners insensitive to the prosodic boundary cues present in the signal. Our results provide additional evidence for the Late Closure strategy being favored in syntactic parsing (Frazier, 1979). For ambiguous sentences, parsers attach the new information to the clause being processed. The noun phrases in LC and EC conditions would therefore more likely be attached to the preceding verb and interpreted as direct objects.

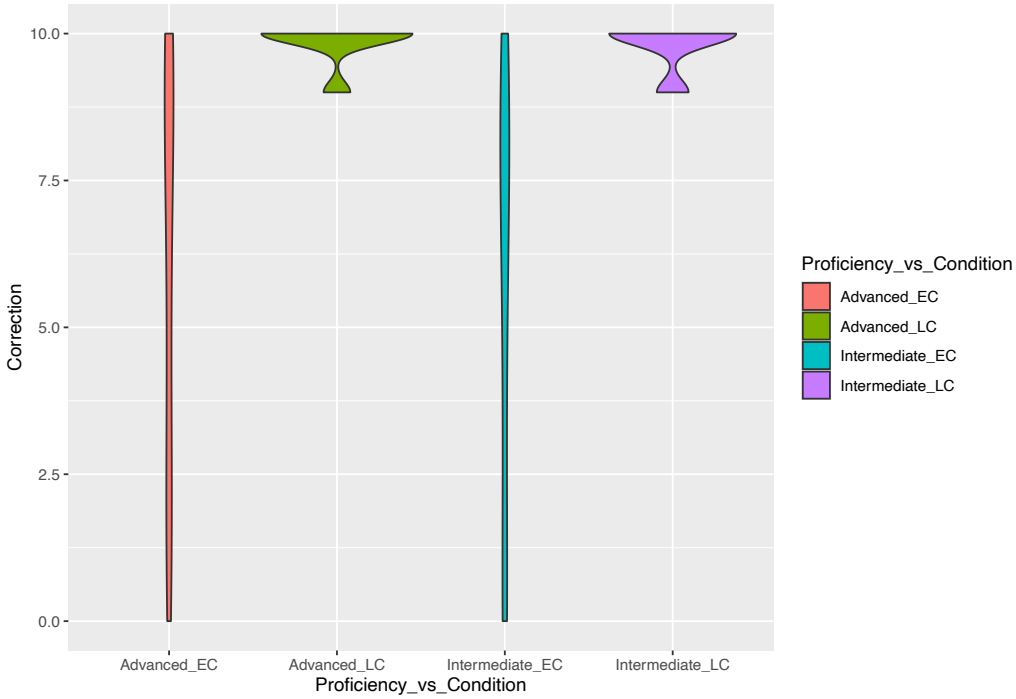


Figure 1: Rate of correction given by Chinese L2 learners for conditions EC vs LC

References

Carlson, K., Frazier, L., Clifton JR., C. (2009a). How prosody constrains comprehension: A limited effect of prosodic packaging. Lingua 119(7), 1066-1082.

Carlson, K., Clifton JR., C., Frazier, L. (2009b). Nonlocal effects of prosodic boundaries. Memory & Cognition 37(7), 1014-1025.

Clifton JR., C., Carlson, K., Frazier, L. (2002). Informative Prosodic Boundaries. Language and Speech 45(2), 87-144.

Frazier, L. (1979). On comprehending sentences: Syntactic parsing strategies (thèse de doctorat, University of Connecticut).

Carryover V-to-V coarticulation in French across different boundaries

Alice Yildiz, Laboratoire de Phonétique et Phonologie, UMR7018, CNRS/Sorbonne Nouvelle
alice.yildiz@sorbonne-nouvelle.fr

Coarticulation is a phenomenon of phonetic influence of a speech unit by other surrounding units. V-to-V coarticulation is the influence of a vowel on another vowel. In a carryover V-to-V coarticulation, the second vowel then takes on some characteristics of the first vowel. D'Alessandro et al. (2022) show that there is anticipatory coarticulation in French affected differently by the strength/size of prosodic boundaries. In particular, they show that there is more coarticulation within a word than between two words, but this would depend on the speakers. In this study we want to shed light on carryover V-to-V coarticulation in French across different boundaries. Our hypothesis is that there would tighter (or stronger) coordination and thus more coarticulation within a word than between two words: i.e. within a word < between two words < between two clauses.

We analyse a corpus of three sentences with the sequence /ipa/ in three different prosodic positions: within a word (1a), between two words in a clause (1b), and between two clauses (1c); repeated 45 times each in 5 sessions, and three control sentences with the sequence /apa/ (2a-c). The sentences were read by 5 native French speakers without any particular instructions.

(1) a. « Quand P**ipa** s'en va au stade, elle râle » /api/
'When Pipa go to the stadium, she complains'
b. « Pap**i** **pa**sse par chez Louis en limousine » /a#pi/
'Grandpa passes by Louis' house in a limousine'
c. « Pap**i**, **pa**ssant par Poitier, l'a vu » /a##pi/
'Grandpa, passing through Poitier, saw him/her/it'

(2) a. « Quand p**apa** s'en va au stade, il râle » /apa/
'When dad go to the stadium, he complains'
b. « Pap**a** **pa**sse par chez Louis en limousine » /a#pa/
'Dad passes by Louis' house in a limousine'
c. « Pap**a**, **pa**ssant par Poitier, l'a vu » /a##pa/
'Dad, passing through Poitier, saw him/her/it'

To measure coarticulation, we took the F1 and the F2 measurements of the two vowels taken at 3 points which we have averaged: 50-60-70% of the first vowel, and 30-40-50% of the second vowel. To measure the effect of the first vowel on the second vowel, we (1) took the distance of F1 and F2 of the second vowel /a/ in the sequences /ipa/ and /apa/, and to measure the degree of carryover V-to-V coarticulation, we (2) used the coarticulation index as in D'Alessandro & Fougeron (2021) and D'Alessandro et al. (2022):

$$Coarticulation\ index = \frac{(F1 - F2)_{V2} - (F1 - F2)_{V1}}{(F1 - F2)_{V1}}$$

The distance between F1 and F2 of the second vowel /a/ in /ipa/ compared to the /apa/ control sequence show that there is less compactness in the /ipa/ sequence meaning an influence of the first vowel /i/ on the second vowel /a/ (Figure 1), but the degree of influence differs according to the prosodic positions.

When looking at prosodic boundaries, contrary to our hypotheses, the /ipa/ sequence with the least coarticulation was the sequence within a word, while the sequence between two words showed most coarticulation (Figure 2). In French, the accent is placed at the end of the accentual phrase (Jun & Fougeron, 2002), and studies show that prosody and stressed vowels can affect coarticulation (Cho, 2004; Recasens, 2015), the stress of the target vowel can make it more resistant to coarticulation (Conklin & Dmitrieva, 2019).

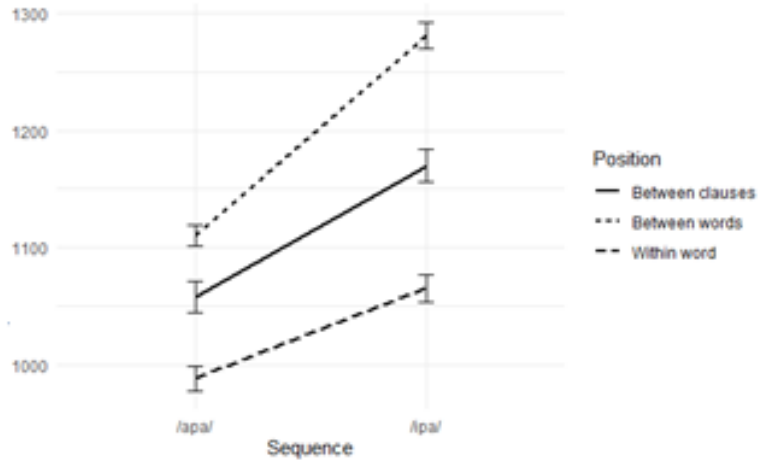


Figure 1: F1-F2 distance of the second vowel /a/ in /ipa/ and /apa/ across different boundaries.

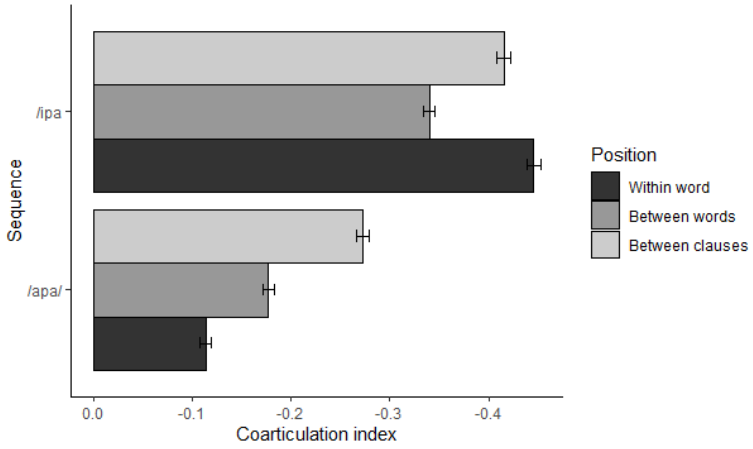


Figure 2: Coarticulation index of the second vowel /a/ in /ipa/ and /apa/ across different boundaries.

References

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 32(2), 141-176.

Conklin, J., & Dmitrieva, O. (2019). Vowel-to-Vowel Coarticulation in Spanish Nonwords. *Phonetica*, 77(4), 294-319.

D'Alessandro, D., & Fougeron, C. (2021). Changes in Anticipatory VtoV Coarticulation in French during Adulthood. *Languages*, 6(4), 181.

D'Alessandro, D., Yildiz, A., Fougeron, C. (2022). Variation individuelle de la coarticulation en fonction de la frontière prosodique. In *Journées d'Etudes sur la Parole 2022* (JEPs 2022), Île de Noirmoutier, France.

Jun, S. A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus*, 14(1), 147-172.

Recasens, D. (2015). The Effect of Stress and Speech Rate on Vowel Coarticulation in Catalan Vowel–Consonant–Vowel Sequences. *Journal of Speech, Language, and Hearing Research*, 58(5), 1407-1424.

Are speakers' formant frequencies predicted from their height and weight?: An acoustic study

Dayeon Yoon, Laboratoire de Phonétique et Phonologie (CNRS/Univ. Sorbonne Nouvelle, France)
dayeon.yoon@sorbonne-nouvelle.fr

Human vocal tract length (VTL) is predicted to be longer in males than in females, which leads to lower formants frequency values in males (Fant, 1970). Based on the hypothesis that VTL must not be free from overall body size (Fitch, 1997; Fitch and Giedd, 1999), many researchers have explored the relationship between body size and formant frequencies, but results have not reached a consensus. While Dusan (2005), Fitch (1997), Fitch and Giedd (1999), and Johnson (2006) found a good correlation between body size and formant frequencies, Gonzalez (2004) and Van Dommelen and Moxness (1995) did not. Meanwhile, it has been reported that the relationship could be sex-dependent (Fitch and Giedd, 1999; Van Dommelen and Moxness, 1995). In line with this previous work, our objective in this study is to evaluate whether physical attributes may account for acoustic properties of speakers. We also aim to investigate if the association between speakers' body size and formant frequencies would differ depending on the speakers' sex.

In our experiment, 33 male and 35 female speakers aged from 18 to 42 participated in the study. Height and weight were self-reported (mean height of 176.39 cm +/- 6.24 cm in males, 161.97 cm +/- 5.57 cm in females; mean weight of 71.75 kg +/- 10.87 kg in males, 54.62 +/- 6.93 kg in females). To acoustically estimate the maximum articulatory space, the speakers produced several "diphthongs": vocalization of a maximally closed [i]-like vowel and a maximally open [a]-like vowel (jaw-opening movement); a mid vowel with a progressive tongue retraction from the most anterior (tongue tip between the teeth, for example) to posterior part of the oral cavity (front-to-back tongue movement); a mid vowel with lips maximally retracted first and then maximally rounded with protrusion (lip-retracting/rounding movement). Each task was repeated three times. F1, F2, F3 values were then measured with Praat at the beginning and the end of the sound production. Euclidean distance (ED) from the beginning to end of the "diphthong" was computed in the F1/F2/F3 space: $ED_{(Fi/Fj/Fk)} = \sqrt{((F_{i\_beg} - F_{i\_end})^2 + (F_{j\_beg} - F_{j\_end})^2 + (F_{k\_beg} - F_{k\_end})^2)}$. A linear mixed effect model was used to test the effect of sex and body size on the $ED_{(F1/F2/F3)}$ ($ED_{(F1/F2/F3)}$ ~ sex*height (or, sex*weight) + (1|speaker)), and Pearson correlations were computed between the ED, individual formant frequency values, and body size.

Our results show that speakers' height and weight could be a cue for their formant frequencies. This tendency is more clearly found in the acoustic distance from the maximally high to low vowels: during the vertical excursion of jaw, the ED was shorter as the body height and weight increased. For the front-to-back tongue movement, an interaction was found between male and female speakers. During the lip-retracting/rounding movement, no significant relationship was found between the speakers' body size indices and the acoustic distance. Regarding the effect of sex, there was no difference between males and females in the relationship between the acoustic distance and the body size in all tasks. The specific data and the findings of this study will be discussed in more detail during the summer school.

References

Dusan, S. (2005). Estimation of speaker's height and vocal tract length from speech signal. In *Proceedings of the 9th European conference on speech communication and technology (Interspeech 2005)*, Lisbon, Portugal, 4-8 September, 1989-1992.

Fant, G. (1970). *Acoustic Theory of Speech Production*. The Hague: Mouton De Gruyter.

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America, 102,* 1213-1222.

Fitch, W. T. & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106,* 1511-1522.

Gonzalez, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, *32*, 277-287.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485-499.

Van Dommelen, W. & Moxness, B. (1995). Acoustic Parameters in Speaker Height and Weight Identification: Sex-Specific Behaviour. *Language and Speech, 38*, 267-287.

Glottal Inverse Filtering Based On Articulatory Synthesis And Deep Learning

Xinyu Zhang, Institute of Acoustics and Speech Communication, TU Dresden
xinyu.zhang1@tu-dresden.de

As the excitation signal for human pronunciation, the glottal airflow carries a variety of information about the speaker's anatomy, the phonation type, and the movement of the vocal folds, which is significant for the speaker identification and emotion recognition. However, the position of the glottis makes it difficult to use an invasive and effective method to measure the glottal airflow signal.

Therefore, we proposed a new method to estimate the glottal vocal tract excitation from speech signals based on deep learning. A BiLSTM neural network was trained to predict the glottal airflow derivative from the speech signal. We used the articulatory speech synthesizer VocalTractLab (VTL) to generate a large dataset containing synchronous connected speech and glottal airflow signals for training. The trained model's performance was evaluated by means of stationary synthetic signals from the OPENGLOT glottal inverse filtering benchmark dataset and by using our dataset of synthetic speech. The performance was determined both in terms of the difference of the open quotient ΔOQ and the cross-correlation r between the ground truth $u_g(k)$ and predicted glottal flow $\hat{u}_g(k)$. The results of the best-performing model are shown in Table 1 using IAIF as a reference. A representative sample of the estimated glottal flow signals in OPENGLOT is shown in Figure 1.

Table 1: Performance of the proposed BiLSTM model and the reference via IAIF. Results are shown as mean ± std (median).

	Proposed BiLSTM		IAIF	
	$\lvert\Delta OQ\rvert$	$r(u_g(k), \hat{u}_g(k))$	$\lvert\Delta OQ\rvert$	$r(u_g(k), \hat{u}_g(k))$
VTL test subset /a, e, i, o, u, ɛ/	**0.0291±0.0177(0.0269)**	**0.9501±0.0370(0.9612)**	0.0879±0.0489(0.0765)	0.8451±0.0516(0.8475)
OPENGLOT				
Repository I	0.0461±0.0461(0.0361)	0.9128±0.0785(0.9413)	**0.0227±0.0627(0.0105)**	**0.9756±0.0174(0.9801)**
Repository II	0.0671±0.0684(0.0359)	0.9167±0.0608(0.9347)	**0.0206±0.0164(0.0171)**	**0.9782±0.0204(0.9878)**
Repository III	**0.1937±0.0953(0.1758)**	**0.8578±0.1010(0.8854)**	0.2166±0.0421(0.2029)	0.8657±0.0304(0.8683)

Compared to the state of the art, the performance of the BiLSTM was slightly worse for repository I and II of OPENGLOT and slightly better in repository III. However, our model was much more accurate and plausible on the connected speech signals, especially for sounds with mixed excitation (e.g. fricatives) or sounds with pronounced zeros in their transfer function (e.g. nasals). As shown in Figure 2a, the signal estimated by IAIF was much less accurate compared to the proposed model in such cases. Furthermore, a qualitative analysis of the glottal flow estimations for the natural speech signals from the BITS corpus was performed (Figure 2b). As the results of the synthetic dataset, the output of the proposed model is more consistent with the suggestions of EGG data.

In addition, IAIF requires the manual specification of some parameters based on the speech signal content, while the BiLSTM model has no free parameters and can process continuous, arbitrary speech input including voiced/unvoiced transitions without any user intervention.

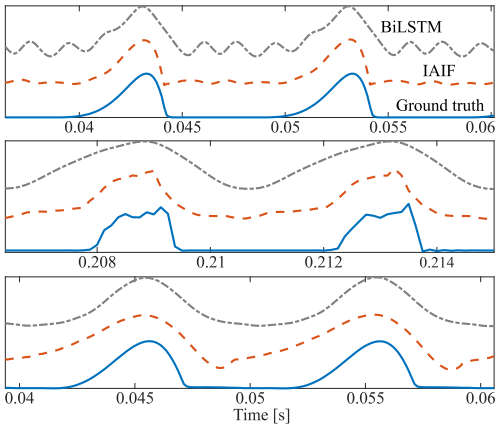
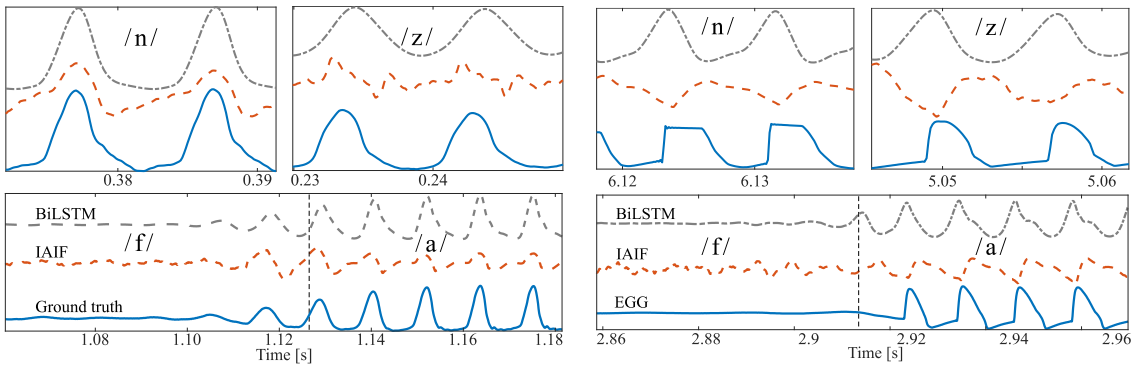


Figure 1: Examples of glottal flow estimations (top to bottom): average error in repository I, highest error in repository II, and lowest error in repository III.



(a) (b)

Figure 2: (a) Example glottal flow segments for the phonemes /n, z/ and an unvoiced/voiced transition from the VTL test set of synthetic speech.

(b) Example glottal flow and EGG segments for the phonemes /n, z/ and an unvoiced/voiced transition from the BITS corpus of natural speech.

References

Alku, P. (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, *11*(2-3), 109-118.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PloS one*, *8*(4), e60603.

Alku, P., Murtola, T., Malinen, J., Kuortti, J., Story, B., Airaksinen, M., ... & Geneid, A. (2019). OPENGLOT–An open environment for the evaluation of glottal inverse filtering. *Speech Communication*, *107*, 38-47.

Timcke, R., von Leden, H., & Moore, P. (1958). Laryngeal vibrations: Measurements of the glottic wave: Part I. The normal vibratory cycle. *AMA Archives of Otolaryngology*, *68*(1), 1-19.

Ellbogen, T., Schiel, F., & Steffen, A. (2004). The BITS speech synthesis corpus for German. *age*, *47*(45), 40.

The Impact of Schwa Deletion on Perceived Tempo in English

Leendert Plug [1], Yue Zheng [2], Robert Lennon [3], Rachel Smith [4]
[1,2] University of Leeds, [3,4] University of Glasgow
[2] y.zheng@leeds.ac.uk

The study reports on experiments aimed to test the hypothesis that English listeners orient to canonical forms in judging the tempo of reduced speech. To date, few studies have assessed how listeners estimate the tempo of speech that features deletions. Orientation to canonical forms should mean that perceived tempo is higher than if listeners orient to articulation rate calculated based on surface phone strings. However, some researchers (Koreman, 2006; Reinisch, 2016) have suggested this may be explained by listeners' learned association between fast speech and phonetic reduction. In the experiments reported here we therefore minimised variation in speech style. In English, the non-realisation of schwa in an unstressed syllable (e.g. support) may result in a surface consonant cluster associated with a different word than the intended one (sport). We presented listeners with sentences containing such ambiguous surface realisations, along with written versions of the sentences to convince some that they were listening to disyllabic words (support etc.) and others that they were listening to monosyllabic ones (sport etc.). Asking listeners to judge the tempo of the sentences allowed us to assess whether the difference in interpretation had an impact on perceived tempo. The results reveal the predicted effect of the imposed word interpretation: sentences with an imposed 'disyllabic' interpretation for the ambiguous word form were judged faster than (the same) sentences with an imposed 'monosyllabic' interpretation. The study also reveals an effect of the order of sentence, which should be accounted for in future studies.

References

Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. Journal of the Acoustical Society of America, 119, 582-596.

Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. Attention, Perception & Psychophysics, 9, 9.