SimPhon.Net

together with

Stuttgart Research Focus (SRF) Language and Cognition

workshop 5

Psycholinguistic, cognitive and neurolinguistic modeling in phonetics and phonology

June 4–6, 2018

Abstracts

SimPhon.Net is a network of close interdisciplinary collaboration between linguists and computer scientists. It addresses the challenge to model and simulate phonetic variability. Through experiments with computer simulations we can pose a variety of questions to unobservable or inseparable aspects of phonetic processes and phonological systems.

The focus of this workshop is on *Psycholinguistic, cognitive and neurolinguistic modeling in phonetics and phonology.*

This workshop is funded by the German Research Foundation (DFG); together with the Stuttgart Research Focus (SRF) Language and Cognition.

Organizers:
Natalie Lewandowski, Daniel Duran and the members of SimPhon.Net and SRF.

Venue:
The workshop is hosted by SimPhon.Net at the *Christkönigshaus* [www.christkoenigshaus.de].

http://www.simphon.net/workshops.html

Abstracts

Invited talk: "Executive functions and L2 phonological development: a study in multiplicity"

Isabelle Darcy
Department of second language studies
Cognitive science program
Indiana University

Executive functions, such as attention control, selective attention, inhibition, and working memory, are involved in almost every aspect of daily life, including having a conversation with someone. In recent years, research has uncovered a multitude of complex relationships between executive functions and language learning. One particularly exciting avenue for this research is to explore the potential relationship between executive functions and phonological learning (in a broad sense) in a second language. For example, the accuracy with which someone is able to pronounce their second language could be related to their individual ability to pay attention to some of its phonetic characteristics. In this presentation, I will synthetize findings – obtained both in and outside my lab – showing that executive functions predict phonological processing and phonological learning in a second language. I will present data from tasks targeting perception, production and lexical access, which suggest that executive functions may have a profound impact on L2 phonological development, but are as of yet insufficiently understood. In particular, causality is hard to establish. I will also highlight the multiplicity of these relationships as they interact with tasks, with learning context and language use patterns, among others, and outline possible avenues for research on pronunciation instruction.

Invited talk: "Effects of cognitive load on speech perception"

Sven Mattys
University of York

Improving the validity of speech-recognition models requires an understanding of how speech is processed in everyday life. Unlike listening conditions leading to a degradation of the signal (e.g., noise), adverse conditions that do not alter the integrity of the signal (e.g., cognitive load, CL) have been under-studied. Drawing upon behavioural and imaging methods, our research shows that CL reduces sensitivity to phonetic detail and increases reliance on lexical knowledge. Critically, however, we found that increased reliance on lexical knowledge under CL is a cascaded effect of impoverished phonetic processing, not a direct consequence of CL. Findings of increased auditory thresholds under CL add further support to the case for an early locus of interference. The results not only constrain our understating of the functional architecture of speech-recognition models, they also invite a re-analysis of the validity of hearing tests for assessing everyday listening.

Invited talk: "Contributions of auditory attention control to L2 phonetic learning"

Joan C. Mora
Universitat de Barcelona

Well-established sources of inter-learner variability in second language pronunciation include age and experience-related factors such as the age of onset of L2 learning, amount of L2 and L1 use and the quality and quantity of L2 input received. Recent research has contributed significantly to our understanding of L2 phonological acquisition by investigating the role of individual differences in cognitive ability (memory, attention and inhibition) as potential sources of variability. This research has established a positive relationship between cognitive skills and L2 phonological processing skills for both immersion and classroom L2 learners. However, the extent to which gains in L2 phonological development can directly be attributed to a specific cognitive skill is still an empirical question, and at present largely under-researched. In this talk I will present data from L2-English learners who were tested on their auditory attention control skills and trained on the perception and production of difficult L2 vowel contrasts. Learners' individual differences in three components of auditory attention control (selective attention, attention switching and inhibition) were assessed and related to their individual gains in perceptual and productive sensitivity to the target vowel contrasts. We will discuss the results in relation to the role of attention skills in L2 phonetic learning and pronunciation development and propose ideas for the design of pronunciation tasks that integrate attention to phonetic form in L2 pronunciation teaching.

Invited talk: "Predictions during speech sound processing: The role of linguistic and extra-linguistic factors"

Mathias Scharinger
Philipps Universität Marburg

Predictions during language comprehensions are currently discussed from many points of view. Opinions differ between views asking whether predictions are needed at all, and approaches where predictions are key mechanisms of processing. I will provide some evidence from linguistic (phoneme, word, and sentence) and extra-linguistic (speaker, aesthetic evaluation) processing levels that predictions indeed play a role during speech perception. I will try to accommodate these findings into a neuro-biologically grounded speech processing model.

Regular talks

"Experiments with radar-based silent phoneme recognition"

Peter Birkholz
TU Dresden

There is currently an increasing interest in the recognition of silent speech, which has a range of novel applications. A major obstacle for a wide spread of silent-speech technology is the lack of measurement methods for speech movements that are convenient, non-invasive, portable, and robust at the same time. Therefore, as an alternative to established methods, we examined to what extent different phonemes can be discriminated from the electromagnetic transmission and reflection properties of the vocal tract. To this end, we attached two Vivaldi antennas on the cheek and below the chin of two subjects. While the subjects produced 25 phonemes in multiple phonetic contexts each, we measured the electromagnetic transmission spectra from one antenna to the other, and the reflection spectra for each antenna (radar), in a frequency band from 2-12 GHz. Here we report the results of preliminary classification experiments based on the setup using different classifiers and feature vectors.

"A socio-cognitive exemplar model of phonetic convergence"

Daniel Duran [1] *& Natalie Lewandowski* [2]
[1] Albert-Ludwigs-Universität Freiburg; [2] Universität Stuttgart

I present an outline of a socio-cognitive exemplar model of phonetic convergence – a phenomenon whereby interlocutors become more similar to each other's speech within a dialog. Previously, there have been two major explanations (1) the socio-linguistic *Communication Accommodation Theory* according to which convergence is based on the needs of speakers to express their attitudes towards their interlocutors, and (2) a "mechanistic" model (most prominantly by Pickering and Garrod), which is based on the idea of a speech production–perception loop that automatically leads to convergence. A literature review shows that empirical evidence supports both approaches. Convergence is thus attributed to two factors which a complete model of the phenomenon needs to integrate: (1) social processes influenced by speakers' intentions, goals and knowledge and (2) psycholinguistic and cognitive processes linking perception and production. I present a hybrid socio-cognitive exemplar model of convergence which takes into account social factors as well as individual differences of the speakers' linguistic experiences, psychological aspects and also their cognitive processing skills. The model assumes an exemplar-theoretic speech production-perception loop shaping new productions on a collection of previously encountered speech items stored in memory. Taking into account recency, exemplars just heard from an interlocutor may serve as or influence one's own productions. This, however, may be enhanced or hampered by a person's personality trait combination and their respective attention skills. In order to develop a computational model, we created two large databases of German spontaneous dialogs (GECO and GECO 2) with high-quality speech recordings accompanied by personality and cognitive data for all participants. This is based on joint work with Natalie Lewandowski and Antje Schweitzer.

"Phonetic Accommodation in HCI - Planning a WoZ Experiment"

Iona Gessinger[1,2], *Eran Raveh*[1,2], *Bernd Möbius*[1] *& Ingmar Steiner*[1–3]

[1] Language Science and Technology, Saarland University; [2] Multimodal Computing and Interaction, Saarland University; [3] German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken

We are presenting a Wizard of Oz (WoZ) experiment which is currently in its planning stage. The experiment is designed to investigate phonetic accommodation of users in the context of human-computer interaction (HCI). In the experiment, an intelligent spoken-dialogue system (SDS) is simulated, since current SDSs are not yet phonetically responsive to the user input.

With this experiment we contribute to the small set of dialogue studies which have been carried out to examine phonetic accommodation of users in the context of HCI [e.g., 1, 4]. Prior own experimentation has shown that humans exhibit phonetically accommodative behavior on the level of segmental pronunciation, word-level spectral composition, and realization of pitch accents when shadowing synthetic stimuli [2, 3]. This confirms that the phenomenon is not restricted to human-human interaction.

The present WoZ experiment entails a dynamic exchange between system and user which is couched in a language learning scenario with target language German. The dialogue is task-oriented and consists of several sub-tasks during each of which the wizard voice systematically varies certain phonetic features, such as intonation pattern, pitch accent placement, segmental pronunciation, and speaking rate. The tasks will be presented to German native speakers as well as to leaners of German as a foreign language. Participants are expected to accommodate to the system to different degrees depending on their L1, various personality traits, and overall phonetic talent.

[1] Bell, L., Gustafson, J., and Heldner, M. (2003). Prosodic Adaptation in Human-Computer Interaction. In *ICPHS*, pages 833–836, Barcelona.

[2] Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., and Steiner, I. (2017). Shadowing Synthesized Speech – Segmental Analysis of Phonetic Convergence. In *Interspeech*, pages 3797–3801, Stockholm.

[3] Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., and Steiner, I. (2018). Convergence of Pitch Accents in a Shadowing Task. In *Speech Prosody*, Poznań, Poland. [in press].

[4] Oviatt, S., Darves, C., and Coulston, R. (2004). Toward adaptive conversational interfaces: Modeling speech convergence with animated personas. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 11(3):300–328.

"Empirical dynamical modeling of speech and motion data: achievements, perspectives, issues and solutions"

Leonardo Lancia
Laboratoire de Phonetique et Phonologie (CNRS/Sorbonne Nouvelle)

The aim of this talk is to describe an integrated framework conceived to model speech communication as emerging from many heterogeneous processes that interact within and between speakers. The observed quantities are modelled as governed by dynamical systems and the analysis aims at characterizing the web of coupling relations connecting their functioning. The framework is adapted from the Empirical Dynamical Modelling framework introduced by Ye et al. (2015) to study the relations between environment and population biology. The approach is applied to simultaneously observed time series reflecting different aspects of the system under study (e.g. the speech production apparatus of a single speaker or the dyadic system composed by the sensorimotor systems of two speaker in interaction). By quantifying the mutual dependencies between the observed quantities it is possible to build a complex network whose features reflect properties of the system under study. In order to apply this approach, two preliminary steps need to be performed. First, the observed time series must be processed to reflect the properties of the underlying systems. Second, the dependencies between the reconstructed trajectories are quantified. In the presentation I will discuss how these two steps can be adapted in order to deal with the complexities of intentional behaviour in general and of speech processing in particular. I will present the results of the application of these methods to the analysis of several kinds of speech data including amplitude modulation signals, f0 trajectories, articulator movements and breathing cycles.

"MaryTTS: A demonstration of the fully modular TTS system"

Sébastien Le Maguer[1], *Pierre-Antoyne Fontaine & Ingmar Steiner*
[1] Universität des Saarlandes

During the previous workshops, we have presented the evolution of the MaryTTS speech synthesis system. The previous presentations focused on the concepts introduced in this new version and why they are important for research in TTS.

In this presentation, I would like to focus on the actual use and extension of the system. First, I will present the new web interface developped for MaryTTS. Based on this interface, I will explain how to query the system and how to automate the process. Finally, I will explain how to extend and adapt MaryTTS to integrate your research and analyze the TTS process. This explanation will focus on how to embed your process into a MaryTTS module, including which constraints you have to consider and how to configure the process.

"Personality & cognitive factors in phonetic convergence"

Natalie Lewandowski
Universität Stuttgart

Cognitive and psychological factors belong to the rather underresearched aspects bearing influence on phonetic convergence – the adaptation of two speakers' pronunciation styles towards each other. We present data from a conversational laboratory task in a foreign language context (English as an L2) with 20 German divided into two groups according to their phonetic talent, and two English native speakers, who were their dialog partners (Lewandowski, 2012). Convergence was measured objectively, employing amplitude envelopes to determine the degree of change in the speakers' pronunciation similarity from an early to a late stage in the dialogs. The acoustic analysis revealed that the group of talented speakers converged significantly more than the less talented group. The strength of the convergence effect was then related to a range of personality and cognitive features of the L2 speakers obtained in another large-scale study (Dogil & Reiterer, 2009), including the Big Five, Behavior Activation/Inhibition Scale (BIS-BAS scale), working memory (e.g. digit span tests) and attention measures (Simon Test). Factors which proved to significantly impact the degree of convergence in the fitted linear mixed model were openness, neuroticism, Behavior Inhibition score (BIS) and the switch costs in a Simon Test (Lewandowski & Jilka, under review).

"Implementing and Modeling Phonetic Convergence in Spoken Dialogue Systems"

Eran Raveh[1], *Ingmar Steiner, Iona Gessinger & Bernd Möbius*
[1] Universität des Saarlandes

Spoken dialogue systems (SDSs) have become common in communication with computers nowadays: personal assistants in the smartphones, service bots in websites, training and tutoring systems, and more.

Investigating the differences between human-computer interaction (HCI) and human-human interaction (HHI) can contribute to the understanding of how they influence efficiency, naturalness, and the overall experience of the interaction from the user's perspective.

It was found that in HHI the interlocutors' behavior converge on different levels.

Some aspects of this phenomenon were integrated into SDSs, e.g., by matching the system's decisions, expectations, lexical choices, etc. to those of the user.

As speech is the primary modality used in SDSs, making the system's behavior more similar to the human interlocutor is expected to improve the naturalness and efficiency of the interaction.

This talk focuses on phonetic convergence and segmental-level variation in particular, including how it can be modeled, and the technical challenges of integrating convergence capabilities into SDSs.

We present a computational model which is used in a customizable end-to-end system that can be used for conducting experiments about phonetic convergence in HCI.

"How to dodge the question in competitive dialogs"

Uwe D. Reichel[1] *& P. Lendvai*
[1] Ludwig-Maximilians-Universität München

We examined conversational non-cooperation in human-human task-oriented dialogues in English. Our approach was to identify prosodic and semantic patterns that characterize replies to information-seeking dialogue acts in a corpus holding competitive and cooperative scenarios. We found that communicative means for dodging a question include reduced content-providing and a stronger F0 declination trend. Some prosodic markers also indicate that holding back information increases the speaker's cognitive workload. Finally, we present initial results on the automatic detection of non-cooperative replies.

"Exemplar-based detection of prominence categories"

Antje Schweitzer
Universität Stuttgart

This talk explores an exemplar-theoretic approach to the integration of phonetics and phonology in the prosodic domain. In an exemplar-theoretic perspective, pitch accent categories are assumed to correspond to accumulations of similar exemplars in an appropriate perceptual space. It should then be possible, as suggested for instance by Pierrehumbert (2003), to infer the (phonological) prosodic categories by clustering speech data in this (phonetic) space. The present experiments explore this approach on one American English and two German databases. They extend an earlier study (Schweitzer, 2011) by assuming a purely phonetic space with more acoustic-prosodic dimensions than the earlier study, while excluding higher-linguistic or phonological dimensions, and by suggesting a procedure that adjusts the space for clustering by modeling the perceptual relevance of these dimensions relative to each other. The procedure employs linear weights derived from a linear regression model trained to predict categorical distances between prominence categories from phonetic distances using prosodically labeled speech data. It is shown that clusterings obtained after adjusting the perceptual space in this way exhibit a better cluster-to-category correspondence, which is comparable to the one found for vowels, and that both the detection of vowel categories and the detection of prosodic categories benefit from the perceptual adjustment.

"Simultaneous Dynamic 3D Face Scanning and Articulography"

Ingmar Steiner[1], *Fabian Tomaschek, Timo Bolkart*[2], *Alexander Hewer*[1] *& Konstantin Sering*[3]

[1] Universität des Saarlandes & DFKI; [2] MPI-IS Tübingen; [3] Eberhard Karls Universität Tübingen

We present a pilot experiment for simultaneous facial and intra-oral motion capture. Our approach combines a state-of-the-art 4D face scanner (3dMD) with electromagnetic articulography (NDI Wave). The data covers a subset of the TIMIT prompt list read by two male subjects, as well as initial data from a male test subject.

This talk gives an overview of the recording procedure and data processing pipeline, and showcases preliminary results in the form of a 3D talking head.

This project is a collaboration between the Cluster of Excellence "Multimodel Computing and Interaction" at Saarland University, the Quantitative Linguistics group at the University of Tübingen, and the Perceiving Systems group at the Max Planck Institute for Intelligent Systems, Tübingen.

"Automatic segmentation and annotation of strongly reduced words"

Fabian Tomaschek[1], *Sam Tureski, Ryan Callihan*

[1] Eberhard Karls Universität Tübingen

In our Karl-Eberhards-Corpus, we have manually corrected word annotations, but lack those for segments. Rerunning a forced aligner on the basis of those corrected annotations results in roughly 30% of non-annotated words, because the forced aligner we are using does not accept words which are shorter than 150 ms. One solution would be to use more context for these words. This however results in wrongly annotated words because words shorter than 150 ms are usually strongly reduced and lack lots of their canonical phones. The forced aligner wants to map canonical phones on a non-existing signal which results in wrong annotations.

We are therefore developing a two-step process in which want to provide annotations for such words. The aim of the talk is to present the procedure in its current stage and discuss potential improvements.

"Nonword segmentation: A new method for assessing speech processing cross-linguistically"

Laurence White

Newcastle University

Understanding how infant and adult language learners separate the speech stream into words – "speech segmentation" – is a theoretical challenge with technological implications. Current experimental methods often rely on language-specific word knowledge, making it difficult to compare the use of segmentation cues (e.g., pitch/timing) between languages. To address such problems, we are developing a new experimental method: "nonword segmentation". This method tests every listener on the same set of cues, regardless of native language, and will provide reaction times as well as accuracy data, for a fuller picture of relative cue weighting.