SimPhon.Net

workshop 1

April 6–8, 2016

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.

Team:

- Jun.-Prof. Dr.-Ing. Peter Birkholz (TU Dresden)
- Dr. phil. Daniel Duran (Universität Stuttgart)
- Dr. James Kirby (University of Edinburgh)
- Leonardo Lancia (Max Planck Institute for Evolutionary Anthropology)
- Dr. phil. Natalie Lewandowski (Universität Stuttgart)
- Prof. Dr. Bernd Möbius (Universität des Saarlandes)
- Dr. phil. Uwe Reichel (Ludwig-Maximilians-Universität München)
- Elena Safronova (Berlin / Universitat de Barcelona)
- Dr. Ingmar Steiner (Universität des Saarlandes / DFKI)
- Dr. Fabian Tomaschek (Universität Tübingen)
- Prof. Dr. Petra Wagner (Universität Bielefeld)
- Andrew Wedel, PhD (University of Arizona)
- Dr. Laurence White (Plymouth University)
- Dr. phil. Frank Zimmerer (Universität des Saarlandes)

This workshop is aimed at computational approaches to the study on human speech segmentation and acquisition. In addition, contributions related to other topics of general interest to the research network are presented. The workshop is organized by the members of SimPhon.Net, funded by *Deutsche Forschungsgemeinschaft (**DFG**)*.

Organizers:
Daniel Duran, Fabian Tomaschek and the members of SimPhon.Net.

Venue:
The workshop is hosted by SimPhon.Net at the Waldhotel Zollernblick:

Waldhotel Zollernblick
Am Zollernblick 1
72250 Freudenstadt - Lauterbad
Germany
www.zollernblick-lauterbad.de

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

Abstracts

"Modeling the role of durational information in human speech segmentation"

Invited Talk: *Odette Scharenborg (Radboud University Nijmegen)*

Computational modelling has proven to be a valuable approach in developing theories of spoken-word processing. In this talk, I will illustrate this claim by presenting two simulation studies carried out with Fine-Tracker. Fine-Tracker is a computational model of human spoken-word recognition based on the theory of spoken-word recognition that assumes that the spoken-word recognition process consist of two consecutive stages, with an 'abstract' discrete symbolic representation at the interface between the stages. The two studies focus on the role of durational information in resolving temporary ambiguity due to lexical embedding (i.e., 'ham' in the longer word 'hamster') to aid speech segmentation.

"A computational model for cross-situational acquisition of words before mastering native language phonemic system"

Invited Talk: *Okko Räsänen (Aalto University)*

A computational model for cross-situational acquisition of words before mastering native language phonemic system Abstract: Children learn their native language simply by interacting with their language community. This is an extraordinary achievement when we start to consider the challenges involved in early language acquisition. For instance, starting without any a priori linguistic knowledge, infants have to learn to segment words out of continuous acoustic speech and map these words to certain visual objects and actions in the environment, i.e., to their meaning. Typically, these two problems are studied and modeled separately, word segmentation preceding meaning acquisition. In this presentation, I will first provide theoretical justification why these two problems should be addressed as a simultaneous integrated process. This is followed by a description of a computational model that can concurrently learn word segments and their meanings from real speech by utilizing cross-situational contextual cues, thereby representing a cross-modal generalization of the so-called statistical learning mechanism that is central to the current language acquisition research. Finally, I will show how behavior of the model fits to human data on language learning observed in various experimental settings.

Bio

Okko Räsänen was born in Finland in 1984. He received the M.Sc. degree in language technology from the Helsinki University of Technology, Finland, in 2007, D.Sc. (Tech.) degree in language technology from Aalto University, Finland, in 2013. He also holds the Title of Docent in Spoken Language Processing in the School of Electrical Engineering of Aalto University.

He is currently a postdoctoral researcher at the Department of Signal Processing and Acoustics at Aalto University and was previously a visiting researcher at the Language and Cognition Lab of Stanford University, California, in 2015. His research interests include computational modeling of language acquisition, cognitive aspects of language processing, context-aware computing, multimodal data analysis, and speech processing in general. He is a member of the International Speech Communication Association (ISCA) and the Cognitive Science Society.

"Articulatory speech synthesis with VocalTractLab"

Talk: Peter Birkholz (TU Dresden)

In this talk I briefly introduce the VocalTractLab (VTL) system for articulatory speech synthesis and its main components: a 3d geometric model of the vocal tract, an advanced self-oscillating model of the vocal folds, a simulation of aerodynamics and acoustics, and a model for articulatory control. Finally, I present some synthesis examples and highlight research questions and projects for which VTL is used.

"In silico Phonetics"

Poster: Daniel Duran (Universität Stuttgart)

This presentation provides a critical overview of the history of simulation methods in phonetics and phonology. This method offers an in silico approach to model building, i.e. it allows for systematic computational analyses of models and underlying theories. The origins approach can be traced back to early mechanical speech synthesis (e.g. to von Kempelen's, Kratzenstein's or Abbé Mical's speaking machines of the 18th century). From mechanical to analog electronic and early digital simulations, these approaches are characterised by a focus on speech synthesis. In 1971–1972 Liljencrants & Lindblom present their seminal work that marks the beginning of modern in silico methods in phonetics and phonology which goes beyond speech synthesis or simple phonological rule testing.

"Investigating the role of prosodic predictability in prominence perception"

Poster: Sofoklis Kakouros (Aalto University)

Frequency and predictability effects are known to play an important role in models of human language production and comprehension. At the level of individual lexemes, for instance, predictability has been shown to be an indicator of prominence. Therefore, predictability at different levels of analysis in speech might be also informative in modeling prominence. Here we discuss the role of prosodic predictability in prominence perception on the basis of the acoustic prosodic features in speech as well as individual lexemes. The underlying hypothesis is that prominence perception could be driven by deviations in the listeners' predictions of the upcoming prosodic patterns and thereby capturing the attention. We show our existing findings from computational modeling studies on English continuous speech, indicating that predictability is a strong cue for prominence at both the acoustic and lexical level.

"On the extraction of rhythmic patterns without speech segmentation"

Talk: Lenoardo Lancia (BLRI labex & ASLAN labex)

It is well known that listeners judge consistently the similarity between rhythmic patterns of different languages on the basis of the properties of the acoustic signals. Classical accounts of the rhythmic structure are based on the duration of speech units, therefore these results should imply that in some way listeners segment the speech stream. More recently it has been proposed that the rhythmic structure of an utterance can be extracted

on the basis of acoustic energy modulations (cf. Leong, Stone, Turner & Goswami, 2014). This is a particularly appealing hypothesis, first because it implies that speakers do not need to segment the speech signal and second because it can be related to studies showing the sensitivity of cortical oscillations to the amplitude modulations of the speech sounds. However, previous works relating amplitude modulation and speech rhythm either did not analyse cross-linguistic data (e.g.: Leong et al., 2014) or could not provide strong empirical support to this hypothesis (e.g.: Tilsen and Arvaniti, 2013). We analysed the coordination between the amplitude modulations due to syllable production and those due to the production of prosodic prominence in narrations elicited with the Pear Story technique from German, English, Polish, Italian and French speakers (on average 5 speakers per language). In order to provide different conditions of enunciation, the participants were asked to tell the story twice: first online during the projection of the video, and once more offline without the aid of the video. Our results show that the coordination between syllable and stress production is stronger in Germanic languages than in Romance languages, with some interesting and somewhat deviant result for Polish.

"Modeling of native language impact on speech segmentation"

Talk: Uwe Reichel (Research Institute for Linguistics; Hungarian Academy of Sciences)

It has been shown in artificial language learning experiments that the choice of perceptual cues for word segmentation is partly dependent of the listener's mother tongue. The present study addresses this native language bias for English and Italian within a Bayesian classifier framework. The two languages differ in how to mark word boundaries and word stress by phoneme lengthening. The classifiers are trained on automatically transcribed written word bigram collections for both languages to predict word boundaries by language-related vowel and consonant lengthening features. The interpolation weights of the classification models shed light on the relative and language-dependent importance of vowel and consonant lengthening as well as of their word initial or final location. Based on the data provided by Laurence White the models are furthermore evaluated in how well they extract the vocabulary of several artificial language variants.

"Modelling audio-visual integration. Or, how we deconstructed the McGurk Effect"

Talk: Fabian Tomaschek (Universität Tübingen) & Daniel Duran (Universität Stuttgart)

We present on-going work on modelling audio-visual integration. We re-visit the "McGurk effect" (McGurk & MacDonald 1976) where stimuli with mismatches in visual and auditory information are sometimes fused into a percept which is different from either stimulus. Unlike most previous studies on this phenomenon, we take into account acoustic and articulographic data taken from real word contexts.

"Introduction to Naive Discriminative Learning"

Tutorial: Fabian Tomaschek (Universität Tübingen)

Computational models provide today's linguists with the possibility to formalize his or her theoretical assumptions and draw precise predictions about the speaker's and listener's

linguistic behavior. Basing on simply learning algorithms tested repeatedly in animal learning behavior, the Naive Discriminative Learning (NDL) represents such a possibility. In the tutorial, I will introduce NDL, implemented in the statistical software R. The NDL package enables the user to model the learning process based on single events using the learning model of Rescorla-Wagner (1972) or predicting the final state of learning using the equilibrium equations by Danks (2003). NDL calculates weights and activations between a set of discriminative features (cues) and their outcomes which they activate. Different learning environment can be represented by different combinations between cues and outcomes, thus creating different models of linguistic behavior. NDL can be used to predict response times in lexical decision tasks, neural behavior, phonetic productions, dialectal distances etc.

"Acoustic speech learning without phonemes: Identifying words isolated from spontaneous speech as a validation for a discriminative learning model for acoustic speech learning"

Poster: Fabian Tomaschek (Universität Tübingen)

Sound units play an important role in models of auditory perception. The idea is that during perception listeners parse speech by means of phones and auditory words. Here we present a model trained on 20 hours of spontaneous speech which is not based on phone representations. The accuracy of its recognition is human-like. Our model generates correct predictions about the speed and accuracy of human auditory comprehension. At the heart of the model is a two-layer network whose input units represent summaries of changes in acoustic frequency bands, and whose output units are lexical meanings.

"Timing, Prominence, Discourse and Multimodality"

Poster: Petra Wagner (Universität Bielefeld)

In the last few years, we have established several research strands at Bielefeld University: (1) Modeling phenomena of phonetic timing (e.g. Windmann et al. 2015), (2) timing and form of multimodal feedback in discourse (e.g. Inden et al., 2014; Wagner et al., 2013, Włodarczak et al., 2013; Malisz et al., under rev.), (3) modeling prominence in perception and production (Malisz & Wagner 2012; Arnold et al., 2013; Samlowski et al., 2014; Wagner et al., 2015), (4) modeling discourse- or speaking-style related phonetic phenomena (Betz et al. 2015; Wagner 2013; Wagner, Trouvain & Zimmerer, 2015, Wagner & Windmann, 2016; Ćwiek et al., in prep.) and (5) the general relationship between speech and co-speech gestures (Wagner, Malisz & Kopp, 2014, Samlowski & Wagner, 2016). Much work is carried out from a perspective of Human-Machine Interaction (e.g. Inden et al., 2014; Hönemann & Wagner, 2015; Betz et al., 2015; Hönemann & Wagner, subm.). Apart from this applied perspective, we use computational modeling techniques as a method to probe and further develop existing phonological and phonetic theories (Windmann et al., 2015).

References

- Arnold, D., Wagner, P., & Baayen, H. (2013). Using generalized additive models and random forests to model German prosodic prominence. Proceedings of Interspeech 2013, Lyon, France, 272-276.
- Betz, S., Wagner, P., & Schlangen, D. (2015). Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. Proceedings of Interspeech 2015, Dresden, Germany, 2222-2226.

- Ćwiek, A., M. Włodarczak, M. Heldner & P. Wagner (in prep.) Breathing pauses vs. breath holds – phonetic differences of a pragmatic resource.
- Hönemann, A., & Wagner, P. (2015). Adaptive Speech Synthesis in a Cognitive Robotic Service Apartment: An Overview and First Steps Towards Voice Selection. Tagungsband Elektronische Sprachsignalverarbeitung ESSV 2015, 135-142.
- Hönemann, A. & P. Wagner (submitted). Synthesizing Attitudes in German. submitted to Interspeech 2016.
- Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2014). Micro-timing of backchannels in human-robot interaction. Presented at the Timing in Human-Robot Interaction: Workshop in Conjunction with the 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI2014), Bielefeld, Germany.
- Malisz, Z., J. Skubisz, M. Wlodarczak, H. Buschmeier & P. Wagner (under revision). The ALICO corpus: analysing the active listener. Language Resources and Evaluation.
- Malisz, Z., & Wagner, P. (2012). Acoustic-phonetic realisation of Polish syllable prominence: a corpus study. In D. Gibbon, D. Hirst, & N. Campbell (Eds.), Speech and Language Technology: Vol. 14/15. Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem. (pp. 105-114). Poznań, Poland.
- Samlowski, B., Möbius, B., & Wagner, P. (2014). Phonetic Detail in German Syllable Pronunciation: Influences of Prosody and Grammar. Frontiers in Psychology, 5.
- Windmann, A., Simko, J., & Wagner, P. (2015). Optimization-based modeling of speech timing. Speech Communication, 74, 76-92.
- Wagner, P. (2013). (What is) the contribution of phonetics to contemporary speech synthesis (?). Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag, Studientexte zur Sprachkommunikation, Band 68, (pp. 75–81), Dresden: TUD Press.
- Wagner, P., Malisz, Z., Inden, B., & Wachsmuth, I. (2013). Interaction phonology – a temporal co-ordination component enabling representational alignment within a model of communication. In I. Wachsmuth, J. de Ruiter, P. Jaecks, & S. Kopp (Eds.), Advances in Interaction Studies: Vol. 6. Alignment in Communication: Towards a New Theory of Communication (pp. 109-132). Amsterdam: Benjamins.
- Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and Speech in Interaction: An Overview. Speech Communication, 57, 209-232.
- Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. Journal of Phonetics, 48, 1-12.
- Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., et al. (2015). DIFFERENT PARTS OF THE SAME ELEPHANT: A ROADMAP TO DISENTANGLE AND CONNECT DIFFERENT PERSPECTIVES ON PROSODIC PROMINENCE. Proceedings of the 18th International Congress of Phonetic Sciences.
- Wagner, P., & Windmann, A. (2016). Acted and spontaneous conversational prosody – same or different? Proceedings of Speech Prosody 2016.
- Włodarczak, M., Simko, J., Wagner, P., O'Dell, M., Lennes, M., & Nieminen, T. (2013). Finnish rhythmic structure and entrainment in overlapped speech. In E. - L. Asu & P. Lippus (Eds.), Nordic Prosody. Proceedings of the XIth Conference (pp. 421-430). Frankfurt a.M.: Peter Lang.

"Exemplar Dynamics in Phonetic Convergence of Speech Rate"

Poster: Michael Walsh (Universität Stuttgart)

We motivate and test an exemplar-theoretic view of phonetic convergence, in which convergence effects arise because exemplars just perceived in a conversation are stored in a speaker's memory, and used subsequently in speech production. Most exemplar models assume that production targets are established using stored exemplars, taking into account their frequency- and recency-influenced level of activation. Thus, convergence effects are expected to arise because the exemplars just perceived from a partner have

a comparably high activation. However, in the case of frequent exemplars, this effect should be countered by the high frequency of already stored, older exemplars. We test this assumption by examining speech rate convergence in spontaneous speech by female German speakers. We fit two linear mixed models, calculating speech rate on the basis of either infrequent, or frequent, syllables, and predict a speaker's speech rate in a phrase by the partner's speech rate in the preceding phrase. As anticipated, we find a significant main effect indicating convergence only for the infrequent syllables. We also find an unexpected significant interaction of the partner's speech rate and the speaker's assessment of the partner in terms of likeability, indicating divergence, but again, only for the infrequent case.

"Beating the bounds: Are there universal prosodic cues to speech structure?"

Talk: Laurence White (School of Psychology, Plymouth University)

Variation in the pitch, length and loudness of speech sounds can facilitate listeners' segmentation of the speech stream into words and phrases. Whilst languages differ in how prosodic features are organised with respect to linguistic structure, some form-function associations may be universal. Firstly, the valency of interpretation of prosodic features appears consistent: sounds that are longer, louder or higher in pitch are more salient. Furthermore, the "Iambic-Trochaic Law" captures the observation that sounds made salient through higher pitch or greater loudness are interpreted as sequence-initial, whilst lengthened segments are interpreted as sequence-final.

Although phrase-final lengthening of vowels is ubiquitous and has been shown to be a cue to an upcoming boundary, consonants are lengthened word-initially in several prosodically diverse languages. In a series of artificial language learning experiments, with native speakers of English, Hungarian and Italian, we explore the importance of the localisation of such timing cues to speech segmentation. We find cross-linguistic support for a functional division between vowel lengthening and consonant lengthening, proposing that this reflects a predictive perceptual mechanism which exploits sensitivity to ongoing speech rate.