SimPhon.Net

workshop 4

(Modelling) the larynx and the voice

December 5–7, 2017

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 *(Modelling) the larynx and the voice*. The workshop is funded by *Deutsche Forschungsgemeinschaft (**DFG**)*.

Organizers:
Peter Birkholz, Daniel Duran and the members of SimPhon.Net. We also thank especially Barbara Wrann.

Venue:
The workshop is hosted by SimPhon.Net at the *WEST Hotel* [http://www.west-hotel.de].

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

Abstracts

(in alphabetical order)

Invited talk: "Acoustic analysis of the voice source and its applications to speech synthesis"

Joao Cabral
Trinity College Dublin

Accurate estimation of the voice source and vocal tract from the speech signal is a complex and difficult task. Alternatively, robust and simple speech analysis methods that do not perform good source-tract separation are very popular because they are more robust and provide good performance for several applications, including the domains of automatic speech recognition (ASR), Text-To-Speech Synthesis (TTS), speaker identification, affect computing and speech coding. Although glottal source modelling has been successfully applied in these domains, its application to real technology systems is small. A better method to separate the voice source from the vocal tract is needed to open a new road for future improvements. In this talk, I will give an overview of previous work on acoustic glottal source estimation. I will also present my current research interest on its evaluation. The resulting experimental outcomes are expected to give insights on how to improve the voice source analysis.

My main application of glottal source analysis is speech synthesis. In this area, one of the greatest challenges is to synthesise expressive speech. The variety in speech and voice is much greater for expressive speech than that of the synthetic voices currently used to read text with a "neutral" voice. This requires a better modelling of voice characteristics, emotions and other voice effects that depend greatly on the voice source. I will present my work on using glottal source modelling for voice transformation of human neutral speech and for controlling the voice quality parameters of the source in statistical parametric speech synthesis.

Invited talk: "Synthesis of running speech for studying the mechanisms of speech production : the case of fricatives and trills"

Benjamin Elie
LORIA

Articulatory synthesis is a technique consisting in numerically simulating the physical phenomena involved in speech production. The aim is to numerically reproduce a speech signal that contains the observed acoustic features with regards to the actual articulatory and phonatory gestures of the speaker. The talk will show that this tool may be useful for researchers to solve problems related to phonetics, or more generally speaking, related to speech production. Our applications will be the fricatives and the alveolar trills.

For fricatives, recent numerical studies have evidenced three distinct regimes of production that are controled by the glottal abduction. These regimes are characterized by the balance between the contributions of both the voiced and the frication noise sources. Interestingly, the regime of voiced fricatives, i.e. when both sources have similar energy, is a very unstable regime, namely small perturbations of the glottal abduction degree leads to large variations of the acoustic characteristics of the produced fricative.

Experiments on real subjects have also evidenced these numerical observations, which could explain the different articulatory strategies used by the speakers to ensure the voiced/voiceless contrast during the production of fricatives.

The mechanisms of production of alveolar trills have been recently investigated via the use of a self-oscillating model of the tongue tip. The interest is to study the articulatory and phonatory configurations that are required to produce alveolar trills. Using a realistic geometry of the vocal tract, derived from cineMRI data of a real speaker, the simulation framework is used to study the impact of a set of parameters on the characteristic features of the produced alveolar trills. It shows, for instance, that the production of trills is favored when the distance between the equilibrium position of the tongue tip and the hard palate in the alveolar zone is less than 1 mm, but without linguopalatal contact, and when the glottis is fully adducted.

"Fabrication of multilayer synthetic vocal folds"

Falk Gabriel
TU Dresden

Fabrication of mulitlayer synthetic vocal fold made of two composite silicon with human like tissue properties.

"Obstruent-induced F0 perturbations: challenges for laryngeal modelling"

James Kirby
University of Edinburgh

In this talk I present some recent work on obstruent-induced F0 perturbations (CF0) in a number of languages including French, Italian, Thai, and Madurese. F0 appears to be more or less universally raised after voiceless obstruents, and may be lowered during the closure during phonetically voiced obstruents. However, the extent to which F0 is perturbed from its globally specified intonation target appears to vary with prosodic environment, and the mechanistic relationship between articulatory source and acoustic effect is not well understood. I will conclude by asking whether, and if so which, biomechanical models might shed light on these and other outstanding questions.

"Studying speech as a complex system through time-series methods and human-machine interactions"

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

In the last decades the complex systems approach to speech processing had a strong influence on many areas of theoretical and applied research. In this view, behavioural patterns are modelled as produced by the interactions between many potentially non-linear dynamical systems involving highly heterogeneous quantities and evolving over different time scales. Despite many potential advantages, such conception of speech processing presents new challenges related to data analysis and interpretation. Indeed our analysis strategy must permit accounting for the increased complexity underlying the observed data. In this presentation I will propose an integrated analytical and experimental framework to study the coordinative patterns underlying verbal interactions at many levels of analysis. These range from the microscopic levels, where different

articulators coordinate their behaviours during the production of a speech gesture, to more macroscopic levels, where we observe the coordination between the sensorimotor and physiological processes of different speakers during verbal interactions. The proposed approach combines original versions of state-space methods for the analysis of the coupling between dynamical systems, with an experimental set-up featuring an artificial agent that adapts in real time its behaviour to that of a human speaker during simplified verbal interactions. This approach permits the systematic manipulation of the human speaker's behaviour by parameterizing that of the artificial agent. Therefore while achieving the same exploratory freedom of a simulation it permits analysing in an appropriate fashion the complexity of real human behaviour.

As a concrete application of the proposed framework, I will describe an experiment in which we asked 10 human speakers to repeat the tongue twister /topkop/ during 12 seconds without interruptions simultaneously with an artificial agent designed to repeat the same utterance. In some uninterrupted sequences, the agent was parameterized to produce its syllables simultaneously with the syllables produced by the speaker (in-phase coordination). In other sequences, the agent was parameterized to produce its syllables in-between those produced by the speaker (anti-phase coordination). Human speakers spontaneously tend toward in-phase coordination. Therefore, in the former kind of sequences the agent cooperates with the speaker, while in the latter kind of sequences the two partners compete two impose two different coordinative relations. With such an experimental paradigm we could compare the predictions of two different theoretical accounts of how interlocutors coordinate their behaviour during verbal interactions. Accounts of inter-speaker coordination based on internal predictive models propose that speakers are able to coordinate their behaviour with that of their interlocutors because they can predict and anticipate it. In accounts based on the notion of dynamical coupling, coordination does not require prediction of the partner's behaviour because appropriate coupling between the sensorimotor systems of the two speakers can result in anticipatory behaviour. These accounts make different predictions about the relation between the degree of inter-speaker coordination and tendency to imitate the behaviour of the conversational partner during verbal interactions. According to prediction-based accounts, we should expect a tendency toward imitation regardless of the cooperative/competitive nature of the interaction. Accounts based on dynamical coupling predict that imitation is observed only if it directly favours the specific coordinative pattern produced by the interlocutors, i.e. only during cooperative interactions. We analysed the relation between the degree of f0 imitation observed in human speakers and the stability of the coordinative relation between the two human speakers and the artificial agent. A positive correlation between the stability of inter-speaker coordination and the degree of f0 imitation was observed only in cooperative interactions. However, in line with accounts based on prediction, speakers imitate the f0 of the agent regardless of whether this is parameterized to cooperate or to compete with them.

"From objective evaluation to objective analysis: a starting point"

Sébastien Le Maguer
Universität des Saarlandes

A Text-To-Speech (TTS) system is a complex pipeline whose purpose is to produce the speech signal corresponding to a given text. To assess the rendered signal, objective and subjective evaluations are used. The objective evaluation consists of computing signal distances between the synthesis and a signal produced by the original speaker. For the

subjective evaluation, human subjects are grading the signal overall quality or selecting a preferred signal between a set of samples.

As TTS aims to produce signal for human listeners, subjective evaluation, even biased, remains the gold standard methodology. On the opposite side, objective evaluation is mainly has anchors with the state of the art. Even though, objective evaluation is really limited, it deserves more focus in order to get more precise analysis methodologies.

In this talk, I propose to analyze the limits of the current objective evaluation methodology. Then we will see alternative methodologies which helps to deal with some limits. Finally, we will explore ways to use objective evaluation to guide the subjective evaluation.

"Computational prosodic typology and its application to understudied languages"

Uwe D. Reichel & Katalin Mády
Research Institute for Linguistics Hungarian Academy of Sciences, Budapest

This study aims to find acoustic evidence for selected aspects of prosodic typology, namely:

- Rhythm: syllable vs stress timed
- Constituency: presence vs absence of an accentual phrase
- Headedness: left- vs right

Based on computational prosody stylization of Hungarian, English, French, and German data we examined numerous pitch, energy and rhythm features for their appropriateness to distinguish the languages in the above listed dimensions. From these features we created prosodic typology profiles, for which we will give phonetic interpretations. We further trained and tested random forest classifiers on the prosodically categorized data in order to obtain a proof of concept to apply this classification on prosodically understudied languages, namely Estonian and Slovak.

"F0 Modeling for DNN-based TTS Synthesis"

Francesco Tombini & Ingmar Steiner
Universität des Saarlandes / DFKI

For several years, DNN-based text-to-speech synthesis (TTS) has been the mainstream paradigm. But predicting intonation, specifically F0, is constrained by data-driven feature extraction and low-level prediction. In this talk, we present ongoing work in parameterizing the F0 contour, and using these parameters for flexible intonation in DNN-based TTS synthesis.

"Intrinsic direction-dependent velocities of articulators"

Christian Thiele
TU Dresden

In this project the biomechanically caused direction-dependent differences of the velocities of multiple articulators are systematically analysed using dedicated measurement methods and suitable experimental techniques. Our goal is to provide a uniform macroscopic speech-related biomechanical characterization of articulators which will might improve articulatory speech synthesis.

"How learning morpho-phonological relations affects phonetic encoding: Modeling the duration of morphemic and non-morphemic S"

Fabian Tomaschek
Universität Tübingen

According to one of the most influential models on speech production – the theory of lexical access by Levelt, Roelofs, and Meyer (1999) – speech production is a modular, sequential process, in which a word's conceptual, syntactic, and morphological information is encapsulated from the articulation process. Consequently, the theory does not make any predictions about dynamics of articulatory processes. Contrary to that model, an increasing number of studies show that higher-level lexical properties do influence the fine phonetic details of speech and consequently, such a separation cannot be upheld (Drager, 2011; Lee-Kim, Davidson, & Hwang, 2012). Just recently, Plag, Homann, and Kunter (2017) showed that the acoustic duration of phonologically homophonous word final [s] in American English differs in its morphological function. The current paper further investigates the source of the durational differences found by Plag et al. (2017). For this, we investigated [s] durations in the Buckeye Corpus and how they relate to the morphological function of word final [s] (e.g. non-morphemic, clitics (has, is), genitive singular, plural noun, third person singular, etc.). We demonstrate that the differences in duration between different kinds of final [s] in English result from linguistic experience in relation to its morphological function. We used a Naive Discriminative Learning Model (Rescorla, 1988), which is a two-layer learning network that computes the strength of association between cues and outcomes. These weights can be conceptualized as the extent to which a particular form can be expected to be associated with a particular meaning in the mind of the average speaker (e.g. [s] → plural).

We show that the association weights between phonological and morphological functions in the context of a given word form and its diphones as cues and the morphological function as outcomes (cf. Figure 1), are highly predictive of acoustic duration. For this, we derived two measures from the association weights (cf. Arnold, Tomaschek, Sering, Ramscar, & Baayen, 2017). First, activations which represent the amount of support a morphological function receives from [s]. Second, activation diversities which represent the amount of uncertainty in the system about which morphological function should be activated. The more word final [s] supports a morphological function the longer it is articulated ($\beta = 0.76$, t = 5.05, cf. Figure 2, left). Simultaneously, the larger the uncertainty about the morphological function, the shorter [s] is articulated ($\beta =$ -1.13, t = -20.39, Figure 2, right). In other words, durations decrease under uncertainty, and increase under certainty. These results show that fine phonetic detail is affected by how strong the phonetic signal is associated with a certain morphological function. Consequently, they present additional evidence that a strict separation between lexical and post-lexical processes cannot be upheld and have important implications for models of speech production, especially for the phonology-morphology interaction.

References

Arnold, D., Tomaschek, F., Sering, K., Ramscar, M., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. PLOS ONE.

Drager, K. K. (2011). Sociophonetic variation and the lemma. Journal of Phonetics, 39(4), 694–707. Cross-Language Speech Perception and Variations in Linguistics Experience.

Lee-Kim, S.-I., Davidson, L., & Hwang, S. (2012). Morphological effects on the darkness of english intervocalic /l/. Laboratory Phonologya, 4(2), 475–511.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999, February). A theory of lexical access in speech production. The Behavioral and brain sciences, 22(1).

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

Rescorla, R. (1988). Pavlovian conditioning - it's not what you think it is. American Psychologist, 43(3), 151–160.

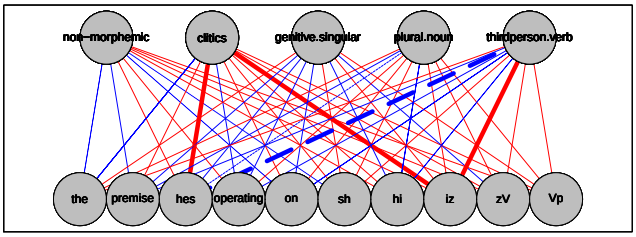

Figure 1. *Cue-to-outcome structure for* the premise he's operating on. *The upper layer represents morphological outcomes, the bottom layer the input, words and their diphones as cues. For diphones, only a subset is presented. Red lines represent positive associations, and blue dashed lines negative associations. Thus, the cue* `he's` *has a positive association strength for the outcome* CLITIC *and a negative association strength with the outcome* THIRD PERSON VERB.

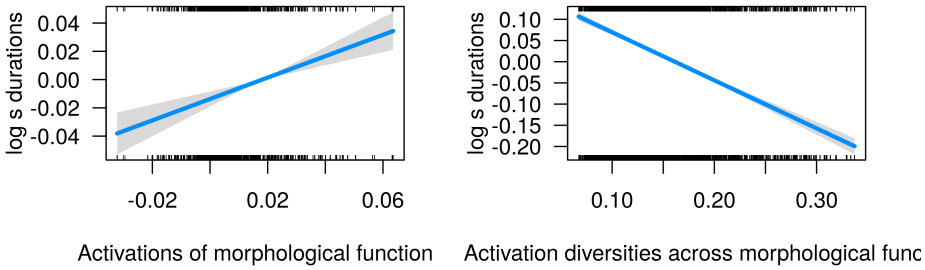

Figure 2. *Partial effects for the duration of homophonous word final [s] in American English. Left panel: Activations from the [s] to the morphological functions. Right panel: Activation diversities for the [s] across the morphological functions.*

"Signal evolution within the word"

Andrew Wedel
University of Arizona

Languages have been shown to optimize their lexicons over time with respect to the amount of signal allocated to words relative to their informativity: words that are on average less predictable in context tend to be longer, while those that are on average more predictable tend to be shorter (Piantadosi et al 2011, cf. Zipf 1935). Further, psycholinguistic research has shown that listeners are able to incrementally process words as they are heard, progressively updating inferences about what word is intended as the phonetic signal unfolds in time. As a consequence, phonetic cues early in the signal for a word are more informative about word-identity because they are less constrained by previous segmental context. This suggests that languages should not only optimize the length of the signal allocated to different words, but optimize the quality and distri-

bution of that information across the word relative to existing competitors in the lexicon. Specifically, words that are on average less predictable in context should evolve more highly informative phonetic cues particularly early in the word, while tending to preserve a longer tail of redundant cues later in the word. In this talk I will review recent work in our group showing that these predictions are borne out in English. I will also briefly present recent statistical work in our group supporting the hypothesis that languages tend to develop phonological grammars which preserve phonological contrasts at the beginnings of words, but reduce contrast later in words. I will argue that this typological tendency plausibly arises from the word-level tendency to preserve higher informativity cues at word beginnings.