Eran Raveh and Iona Gessinger

Saarland University
Computational Linguistics & Phonetics
Saarbrücken, Germany

{raveh|gessinger}@coli.uni-saarland.de

1 Research Interests

Our current research deals with **phonetic convergence** in **human-computer interaction (HCI)**. We are also interested in **intelligent tutoring systems**, especially in the area of **language learning**.

In our current project we are approaching spoken HCI from both the computational and the human perspective, i.e., the user's perspective in the domain of spoken dialogue systems (SDSes).

1.1 Phonetic Convergence

The field of phonetics provides a scientific description for the production of speech. Humans show a high degree of inter-speaker, as well as intra-speaker, variation in their speech productions. In contrast, machines systematically reproduce or reuse speech segments derived from learned patterns using unit selection (Hunt and Black, 1996) or parametric (Zen et al., 2009) speech synthesis, respectively, among others.

In spontaneous conversations, humans tend to phonetically converge to each other, i.e., there is an increase in segmental and suprasegmental similarities between speakers to be observed (Pardo, 2006).

Communication Accommodation Theory (Giles, 1973) postulates that inter-speaker accommodation subserves the function of controlling social distance. Converging to the speech of an interlocutor could thus serve to reduce the social distance.

Apart from the social aspects, phonetic convergence is assumed to support the overall efficiency of spoken interaction by increasing intelligibility and predictability of what is being said.

1.2 Practical Applications

Systems with phonetic awareness and converging capabilities are likely to be more accessible to users than systems that do not accommodate to the speech input. For example, users who are not native speakers of the output language might have a hard time understanding the system's output and would thus benefit from a converging SDS.

Such systems can detect and adapt to variation in the user's speech, potentially making their output more similar to everyday human-human spoken interaction, which may improve the user experience and make the dialogue more fluent and natural.

This kind of adaptation can also be used the other way around, i.e., by *not* converging to the user's speech – perhaps even intentionally diverging from it. This can be effective, for instance, in intelligent tutoring systems for language learning with emphasis on pronunciation, as it is known that pronunciation is usually not emphasized in the language learning process (Grice and Baumann, 2007). Such tutoring systems take advantage of inter-speaker accommodation by giving the learner auditory feedback they can converge to and thus learn to speak in a more native-like manner.

2 Future of Spoken Dialog Research

Multi-system spoken interaction: Nowadays, the most common paradigm for SDSes is a single human interacting with a single machine – e.g., intelligent personal assistance. We believe that one of the next steps is to develop dialogue systems that are also aware of and able to verbally communicate with other machines. This will be handy as soon as speaking machines are more commonly used in everyday interactions and users will be able to dynamically interact with multiple machines to complete a certain task, instead of sequentially completing subtasks with each machine separately. This will also require each system to understand that it was addressed, based on previous utterances and discourse markings.

Tolerance towards speech variation: It is a common phenomenon that users who are not native speakers of American English have more difficulty to be understood by an SDS trained on that particular variety of English. Consequently, those users are forced to adapt their speech to the system. We believe that the system should adapt to the users instead. Therefore, tolerance towards phonetic variation in the speech input needs to be further developed in SDSes. Such tolerance can not only be captured by the learned model, but also by taking language-specific phonetic interferences into account.

Personalized speech output: Personalized speech output already exists in the sense that a device can adapt its underlying model of speech to the user over time. In that way, the speech output of the SDS is refined but does not gain more phonetic flexibility. The above-mentioned tolerance towards speech variation could be translated into linguistic awareness towards speech variation. An SDS that is aware of, and able to produce, phonetic variation could dynamically converge to the speech of different users. This may enhance user experience and make the dialogue more fluent and natural.

3 Suggestions for Discussion

- Objective and subjective evaluation methods for SDSes;
- Development of SDSes for various languages, especially for low-resource languages;
- Improving spoken language understanding (SLU) in SDSes (beyond gap-filling, e.g., using world knowledge or prosodic information).

References

Howard Giles. 1973. Accent mobility: A model and some data. *Anthropological Linguistics* 15(2):87–105.

Martine Grice and Stefan Baumann. 2007. An introduction to intonation – functions and models. In Jürgen Trouvain and Ulrike Gut, editors, *Non-Native Prosody: Phonetic Description and Teaching Practice*, De Gruyter, pages 25–51.

Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. volume 1, pages 373–376.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4):2382–2393.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51(11):1039–1064.

Biographical Sketches

Eran studied Computational Linguistics at the University of Tübingen, Germany, and at Trinity College Dublin, Ireland. His Bachelor thesis introduced a novel approach for automatically generating customizable recall questions for reading comprehension. He deepened his knowledge of speech processing at a startup company in Dublin, developing a multilingual speech synthesis system, and also while working at the University of Stuttgart.

In real life, he plays the piano and saxophone, and occasionally, basketball.

Eran is currently pursuing a PhD in Dr. Ingmar Steiner's group *Multimodal Speech Processing* at Saarland University.

Iona received a BA degree in Romance Linguistics and Phonetics from the University of Jena, Germany, and an MA degree in Speech Science from the University of Marburg, Germany.

She has worked for the German Language Atlas at the University of Marburg and as a teacher for German as a Foreign Language at Lycée Ronsard, Vendôme, France, and at the Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Iona is currently pursuing a PhD with Prof. Bernd Möbius at Saarland University.

Acknowledgments

This research is funded by the German Research Council (DFG) in project "Phonetic Convergence in Human-Computer Interaction (CHIC)".