Paper presented at the 50th Poznań Linguistic Meeting

„Speech-based AI", „digital phonetics" – Speech sciences in the era of machine learning

Zofia Malisz
KTH - Royal Institute of Technology, Stockholm
malisz@kth.se

We are in the middle of a paradigm shift in speech and language technologies. Machine learning methods such as deep neuronal networks have revolutionised the field. For example, deep learning-based speech synthesis is now capable **of imitating human speech so well that it is indistinguishable from natural speech** in behavioural and physiological studies (Malisz et al. 2019a). This rapid progress has consequences for both speech sciences and technology - where new goals and completely new areas of research have begun to emerge.

An essential question remains - how much are systems based on machine learning able to „understand" about speech and human communication? Or rather, **are we able to interpret what these systems are doing** – the same way we can interpret explicit statistical models? Is our expert knowledge, as speech scientists, needed at all anymore for speech technology to develop further? And how can we continue to research and profit from the intersection of speech sciences and technology in a world where a machine "can learn anything"?

I suggest that **there is a lot to do for speech scientists** in the presently emerging new disciplines that may in the future be called: "digital phonetics", "digital linguistics", "speech-based, or "voiced-based AI". Disciplines that works at this intersection.

I point out several current research questions at this intersection that require close co-operation and renewed dialogue between speech engineers and speech scientists:

a) **Explicit control over the details of synthesis output such as duration or pitch.** Speech scientists use synthetic speech in their experiments as stimuli and there, they require a thorough control beyond the current standards in technology. I show that a deep learning -based system can be trained to include explicit control over prosodic prominence, fundamental frequency and formant structures and can be potentially useful for speech scientists (Malisz et al. 2019b, Doehler Beck et al. 2021).
b) Also, **machine learning offers visualisation techniques with human-in-the-loop** that e.g. enable access and fast annotation of speech datasets relevant for research and the general public. I present one such method (Fallgren et al. 2020).

References:

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J., & Gustafson, J. (2019a). Modern speech synthesis for phonetic sciences: A discussion and an evaluation. In 19th International Congress of Phonetic Sciences, Melbourne, Australia.

Malisz, Z., Berthelsen, H., Beskow, J., & Gustafson, J. (2019b). PROMIS: a statistical-parametric speech synthesis system with prominence control via a prominence network. In SSW 10-The 10th ISCA Speech Synthesis Workshop. INTERSPEECH 2019, Graz, satellite workshop.

Döhler Beck,G.T., Wennberg, U., Henter, G.E., Malisz, Z. Wavebender GAN: Controllable speech synthesis for speech sciences. (2021). 1st International Conference on Tone and Intonation 2021, Sonderborg, Denmark.

Fallgren, P., Malisz, Z., & Edlund, J. (2019). How to Annotate 100 Hours in 45 Minutes. In Proceedings of INTERSPEECH 2019, Graz, (pp. 341-345).