Phonetics Colloquium at UdS

on Wednesday, 2 July 2025, 14:15

in building C7.2 in R. 5.09

Nigel Ward

(University of Texas, El Paso, USA)

Implicit and Explicit Prosody Modeling for Speech Synthesis and Dialog Systems

The application of self-supervised learning to massive data sets has revolutionized speech processing, enabling speech synthesizers and spoken dialog systems to often outperform humans. However, they cannot yet manage to sound sincerely positive, enthusiastic, appreciative, context-aware, and so on. This drastically limits their potential uses.

Better modeling of prosody will help. One priority is expanding the usable set of prosodic features, beyond the typical handful, to include voicing properties, phonetic reduction, and others. For example, through tool construction and corpus studies, we could discover that positive feeling is often conveyed in part by phonetic reduction, in both English and Spanish. Another priority is supporting improvements to modern speech synthesizers, whose prosodic behaviors are of course learned rather than hand-crafted. To this end we developed quality metrics that automatically estimate the perceptual distance between any two utterances, in terms of perceived pragmatic functions. These can better quantify the divergence (loss) between system outputs and target utterances, and so enable better training towards objectives.

We are now working to use these findings and tools to improve the efficacy and trustability of an AI co-player for a simple video game.