TTS evaluation seminar at KTH Speech, Music & Hearing

Wednesday 15th June 2022, 14:30 - 16:30 Swedish time

Evaluation of TTS (text-to-speech, speech synthesis) is a much-talked about topic. The new neural frameworks for training TTS deliver voices that in some cases and under some conditions reach levels where TTS is indistinguishable from human speech. An argument could even be made that this has been the case for decades now.

At the same time, TTS evaluation is a topic that has shown comparatively little progress over the decades. Although there is a general awareness in the field that MOS tests are not very reliable yet quite resource consuming, they remain the de facto standard not only in industry but in much of research as well. To be provocative, the greatest changes of these tests since they were developed by in the CCITT in the 70s is that the rather rigorous requirements under which they were validated and verified have been severely relaxed.

As a part of an international and joint effort to come to terms with TTS evaluation, we invite you to a seminar in which we present some of the basics underlying TTS evaluation of different kinds. The seminar centres around three presentations showcasing current efforts in TTS evaluation:

- Erica Cooper on automatic MOS prediction systems
- Christina Tånnander on ARS-based (Audience Response System) evaluation
- Ayushi Pandei on the relation between segmental properties of TTS and its evaluation

The seminar takes place at KTH's division for Speech, Music & Hearing on the afternoon of Wednesday June 15th, in conjunction with the Swedish phonetics meeting Fonetik 2022, and is organised by Språkbanken Tal, the speech section of the Swedish language research infrastructure Nationella språkbanken (2017-00626_VR) and by the Vinnova supported project Deep learning based speech synthesis (2019-02994).

Erica Cooper has a PhD in speech synthesis and recognition from Columbia University. She is Erica coopercurrently a postdoctoral researcher at the Yamagashi Lab at the National Institute of Informatics in Tokyo. Her recent work on large-scale evaluation on TTS systems, as well as automatic MOS prediction systems has led to the VoiceMOS Challenge.

Subjective listening tests are the gold standard for evaluating synthesized speech, but they are time-consuming and costly, and can't easily be integrated into a rapid experimental iteration cycle. Data-driven automatic methods for predicting human opinions of synthesized speech have thus become a topic of interest, but these require large training datasets and tend not to generalize well. In this talk, we will present a large-scale Mean Opinion Score (MOS) dataset, some methods that can improve the generalization ability of MOS predictors, and the outcomes of the first VoiceMOS Challenge, a shared task for automatic MOS prediction.

Christina Tånnander is an industrial PhD student at Speech, Music and Hearing at KTH Royal institute of Technology in Stockholm. She has worked with speech technology for Nordic languages since 1996 and works on speech synthesis for university textbooks and news at the Swedish Agency for Accessible Media (MTM) since 2006.

Most subjective TTS evaluation concerns the ratings of multiple short speech samples, typically a sentence. Although suitable for many evaluation purposes, and widely used for comparing the quality of different synthetic voices, this method doesn't say much about how it is to listen to a specific synthetic voice for a long period of time, for example when listening to books or newspapers. We will briefly discuss the need for suitable evaluation methods for different purposes, and then present a series of experiments using Audience Response System (ARS) for cost-effective evaluation of longer texts.

Ayushi Pandey is a PhD student at Trinity College Dublin. Under the supervision of Naomi Harte, Julie Carson-Berndsen and Sebastien Le Maguer, she explores the segmental properties of synthetic speech, and their relationship to perceived naturalness. Previously, she has worked on phonemic contrast, resource creation and speech recognition for Indian languages.

Segmental properties of Text-To-Speech (TTS) synthesizers have been studied for their influence on various perceived attributes of synthetic speech. However, they have received very limited attention for modern, neural vocoder-based TTS. Secondly, segments have usually not been described as well-defined phonological classes. In this talk, we will first discuss how contrastive properties of phonemes can provide an array of features for a fine-grained analysis of synthetic speech. Then, based on these features, we will present results from a comparative analysis over 4 different TTS techniques.