3 Summary

Relationships between human perceptions and speech features need to be established, which should lead to machines reaching the human performance in the recognition of speaker characteristics. When such systems encounter transmitted instead of clean speech, e.g. in call centers, the speaker characterization accuracy might be impaired by the degradations inserted by communication channels in the speech signal. However, the effects of degraded speech on paralinguistics has not been much investigated in previous research. This project has assessed the influence of transmission channels on speaker likability and personality recognition regarding a) human perceptions and b) automatic predictions.

The "Nautilus Speaker Characterization (NSC) Corpus"[7] [20] has been compiled and made available for non-commercial research at the BAS CLARIN repository. It contains clean conversational speech from 300 German speakers and labels of interpersonal traits (likability, confidence, maturity, etc.) and voice descriptions (bright, articulate, melodious, etc) obtained by subjective listening. These data has served as basis for the subsequent project investigations.

Regarding human speaker characterization, the transmission of voices through wideband (WB, 50–7,000 Hz) instead of narrowband (NB, 300–3,400 Hz) channels greatly influences subjective attributions of speakers' interpersonal characteristics. Male speakers are perceived as more modest, unobtrusive, adult, and sympathetic. If speakers are likable (warm and attractive), then males are also perceived as more decided and females as more competent and beautiful in WB compared to NB.

On the automatic side, different machine learning experiments have addressed the predictive modeling of speaker characteristics, employing clean speech data for model building. State-of-the-art classification techniques have been explored. It has been shown that both encoding and bandwidth limitation artifacts of test speech affect the performance with respect to clean speech sampled at 48 kHz. Given that NB speech causes the classification performance to drop significantly compared to WB and to super-wideband (SWB, 50–14,000 Hz) speech, it can be asserted that the 3,400–7,000 Hz range is crucial for a more accurate classification of voice likability. The influence of different codec schemes and of network transmission errors (packet loss and jitter) have also been examined in detail.

The principal contributions of this work are thus i) the recorded and labeled NSC corpus, ii) the evaluation of transmission bandwidth effects on subjective speaker attributions, iii) the assessment of speech degradations and transmissions effects on automatic speaker characterization, and iv) knowledge about the relationships between voice likability on perceived speech quality.

Future work may concentrate on different directions pursuing adaptive spoken dialog systems. First, the dialog turns of the NSC corpus can be used for the analysis of speech production and conversational behavior in human-human interactions. This can assist the design of dialog strategy and language generation mechanisms. Second, research can be conducted towards the link between acoustic speech parameters and voice descriptions, which can easily be applied to the recognition of speaker interpersonal characteristics, as shown in this project. Finally, building on the provided insights on speaker characterization and voice likability related to quality, voice likability measurements can be incorporated to the process of network planning and communication channel design.

[7] Further details in: http://www.qu.tu-berlin.de/?id=nsc-corpus