Technically enabled explaining of voice characteristics

Jana Wiechmann[1], Thomas Glarner[2], Frederik Rautenberg[2], Petra Wagner[1], Reinhold Häb-Umbach[2]

[1]*Faculty of Linguistics and Literary Studies, Bielefeld University*

[2]*Faculty of Electrical Engineering, Informatics and Mathematics, Paderborn University*

Voice is a complex phenomenon which cannot be described easily. Many different descriptive dimensions and terms exist for characterizing a speaker and the corresponding para- and extralinguistic features of their speech, with different descriptive approaches used by experts, non-experts, and several fields (e.g., Kreiman & Sidtis, 2011). Among these are phoneticians, singers, vocal coaches, actors, speech therapists, people in forensics and more, with varying levels of expertise and interests. There is no consistent vocabulary for describing and explaining voices. Furthermore, it is difficult even for 'experts' to perceive and name the complex range of voice characteristics (e.g., Kreiman and Gerrat, 2000).

However, an internal consistency in characterizing voices in professional contexts is crucial: speech therapists need to have a similar understanding to be able to diagnose a voice disorder and to plan a suitable therapy. It therefore is of great importance that the experts who teach them are able to explain the complex phenomenon of voice to novices in a consistent fashion.

Unlike humans, voice conversion systems have been shown to disentangle voices and voice characteristics from their linguistic content. We therefore use a state-of-the-art voice conversion system in a two-fold fashion: (1) to disentangle the complex dimensions of human voices and serve as a diagnostic tool for voice characterization, (2) to support the explanation of voice characteristics, by demonstrating audibly, how a voice changes in one particular dimension of interest.

The architecture of the voice conversion system is based on an autoencoder with two separate encoders: The content encoder, to extract a subsampled time series of encodings representing the linguistic content from an utterance, and the style encoder, to extract a single encoding per-utterance, mostly representing the voice characteristics from the utterance. The system is trained with adversarial loss terms, leading to disentangled representations, i.e., the voice encoding is as non-predictive for the content as possible and vice versa. Voice conversion can happen either by changing specific components of the voice encoding or by replacing it with one extracted from another utterance altogether. The voice conversion was designed to work with English utterances, but since no assumption about a specific language has been made in the model architecture, the training corpus can be readily exchanged with a German corpus.

To establish a baseline of humans' ability to characterize voices we currently collect data of speech therapists in training (20+) and non-professional actors (20+) that either mimic (actors) or identify (speech therapists) the characteristics of 20 voices selected from the *The Nautilus Speaker Characterization Corpus* (Gallardo & Weiss, 2018). We investigate their understanding of these features firstly without and later with an explanation given by a phonetic expert. These explanations are based on an established set of voice characteristics (Laver, 1980), for which we have clear-cut expectations with respect to their articulatory and acoustic-phonetic realizations, and which have been identified to be suitable dimensions of voice quality perceptions by laypersons (Gallardo & Weiss, 2018), and which led to a high interrater reliability in experts before and after a discussion (Cohen's Kappa, before: k=0.62, after: 0.85).

Our ultimate aim is to achieve a higher consistency and performance in the explanation of human voices using a voice conversion system to support experts and non-experts.

doi:10.11576/pundp2022-1038

[1] Fernández Gallardo, L., Weiss, B. (2018). "The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions," *in International Conference on Language Resources and Evaluation (LREC)*.
[2] Kreiman, J., Gerratt, B.R. (2000). Sources of listener disagreement in voice quality assessment, *Journal of the Acoustical Society of America, 108*, 1867 – 1879.
[3] Kreiman, J., Sidtis, D. (2011). Voices and listeners: Toward a model of voice perception. *Acoustics Today 7*, 7-14.
[4] Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

doi:10.11576/pundp2022-1038