Automatically identifying perceptually similar voices for voice parades

Finnian Kelly[1, 2], *Anil Alexander*[1], *Oscar Forth*[1], *Samuel Kent*[1],
Jonas Lindh[3], *Joel Åkesson*[3]

[1]*Research and Development, Oxford Wave Research Ltd, Oxford, United Kingdom.*
[2]*Center for Robust Speech Systems (CRSS), University of Texas at Dallas, U.S.A.*
[3]*Voxalys AB, Gothenburg, Sweden.*
{finnian|anil|oscar|sam}@oxfordwaveresearch.com,{jonas|joel}@voxalys.se

Currently, selecting the most appropriate speakers for a voice parade is a costly, labour intensive process. Utilising an automatic system for speaker selection, under the supervision of the forensic expert, could offer the potential to increase the efficiency and to decrease the subjectivity of this process, while at the same time considering a larger number of candidate voices.

In the entertainment industry, the task of voice casting involves the reproduction of dialogue from a movie or video game from one language to another (Obin et al. 2014). The process of casting a voice actor could be greatly accelerated with an automatic approach to searching for a similar voice.

With these applications in mind, this paper presents a pilot study on using an automatic speaker recognition system to identify similar voices. Previous research on perceived voice similarity has pointed toward pitch, formant information and voice quality as being important cues (Nolan et al, 2011, Zetterholm et al. 2010). In this study, auto-phonetic features encapsulating a range of these attributes are used within the automatic system.

A corpus of 175 speakers was obtained from various online sources. The corpus consists of lapel-microphone recordings in varied acoustic environments and speaking contexts. All recordings are in English, with variation in accent.

The iVOCALISE system (Alexander et al. 2016) was used for the automatic speaker recognition experiments. iVOCALISE supports a range of feature and modeling options; here, F0, F1-F4, semi-tones of F0, along with first derivatives, were used as features in an i-vector PLDA (Probabilistic Linear Discriminant Analysis) framework. Three male and three female target voices were selected randomly from the SITW subset, and a comparison score between their recordings and all others in the database was obtained.

This was followed by a listener experiment to assess the perceptual similarity of the target voices and their closest cohort, as ranked by the output score of iVOCALISE. Each target voice was compared with five candidate voices: the two closest voices, two randomly chosen voices and sample of the same voice from a different recording. Each sample was seven seconds in duration and no samples were repeated. The test was completed by 43 listeners (25 male, 18 female) via an online application. After a training phase of three comparisons, the 30 comparisons were presented in a random order. Listeners were requested to judge the similarity of the two voices on scale of 1—9 (see Figure 1), while aiming to ignore the speaker accents, any non-speech noises, or any of the spoken content.

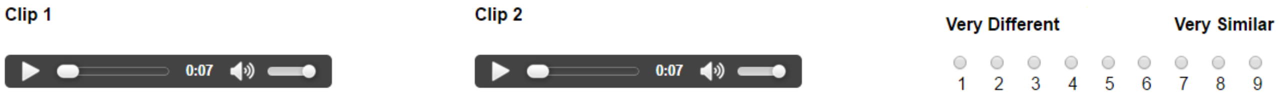

Figure 1. A sample comparison from the listener experiment

Responses for male comparisons are show in Figure 2. A significantly higher response value is observed for 'similar' voices automatically selected using iVOCALISE. Additionally, there is a positive correlation (≈0.7) between score and median similarity rating for all male comparisons. In the female case however, a weaker separation of similar and different comparisons was observed. Considering the unconstrained recording conditions and confounding issues of speaker accent and speech content, this approach of using auto-phonetic features to find similar voices shows promise. With an appropriate database containing meta-data such as accent and age, this could indeed prove a quick and effective method of identifying potential cohorts of speakers for voice parades and voice casting.

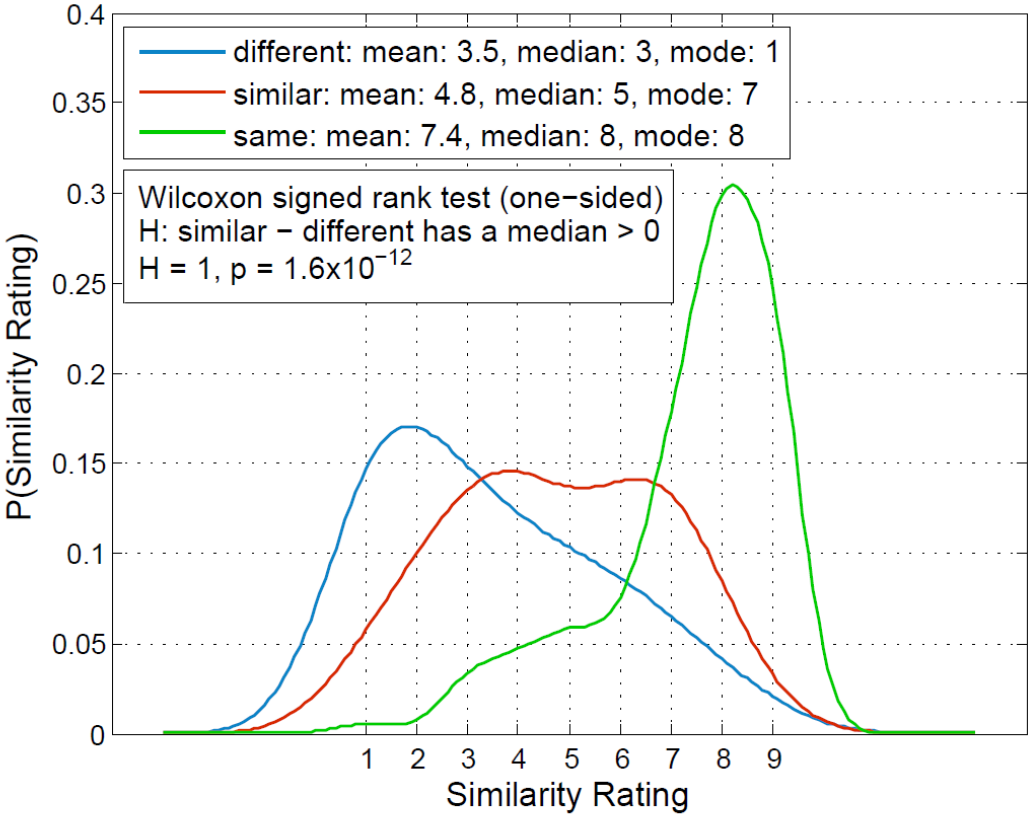

Figure 2. Distribution of similarity ratings for male comparisons from all listeners (15*43 = 645 responses), where 1 indicates 'very different' and 9 'very similar'. The median response for 'similar' comparisons is significantly greater than the median response for 'different' comparisons.

References

Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. (2016). VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features. *To appear at Odyssey 2016, Bilboa, Spain.*

Nolan, F., French, P. McDougall, K., Stevens, L. and Hudson, T. (2011). The role of voice quality 'settings' in perceived voice similarity, *IAFPA conference*, Vienna, Austria.

Obin, N., Roebel, A. and Bachman G. (2014). On automatic voice casting for expressive speech: Speaker recognition vs. speech classification. *In proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, pp. 950-954.

Zetterholm, E., Blomberg, M. and Elenius, D. (2010). A comparison between human perception and a speaker verification system score of a voice imitation. *In proceedings of the 10th Australian International Conference on Speech Science & Technology*, Sydney, pp. 393-397.