INTERSPEECH 2016: Show & Tell Contribution
September 8–12, 2016, San Francisco, USA

Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features

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

Abstract

Assessing the perceptual similarity of voices is necessary for the creation of voice parades, along with media applications such as voice casting. These applications are normally prohibitively expensive to administer, requiring significant amounts of 'expert listening'. The ability to automatically assess voice similarity could benefit these applications by increasing efficiency and reducing subjectivity, while enabling the use of a much larger search space of candidate voices. In this paper, the use of automatically extracted phonetic features within an i-vector speaker recognition system is proposed as a means of identifying cohorts of perceptually similar voices. Features considered include formants (F1-F4), fundamental frequency (F0), semitones of F0, and their derivatives. To demonstrate the viability of this approach, a subset of the Interspeech 2016 special session 'Speakers In The Wild' (SITW) dataset is used in a pilot study comparing subjective listener ratings of similarity with the output of the automatic system. It is observed that the automatic system can locate cohorts of male voices with good perceptual similarity. In addition to these experiments, this proposal will be demonstrated with an application allowing a user to retrieve voices perceptually similar to their own from a large dataset.

Index Terms: voice similarity, voice perception, speaker recognition, phonetic features, speakers in the wild

1. Introduction

Currently, selecting the most appropriate speakers for a voice parade is a costly, labour intensive process. In the U.K., recommendations for creating a voice parade [1] specify that eight comparison voices, or 'foils', should be selected, and together with the suspect's voice, be presented to the 'earwitness'. The candidate voices should be chosen from a pool of at least 20 speakers of 'similar age and ethnic, regional and social background' to the suspect. The selection of voices for the parade is made by a phonetician on the basis of an auditory screening. There have recently been proposals to formalise the selection of a voice parade [2]. However, there still remains significant human effort in this procedure. Utilising an automatic system for the selection of foils, under the supervision of the forensic expert, offers the potential to greatly increase the efficiency of the process, all while reducing subjectivity and considering a candidate voice pool well in excess of the recommended 20 speakers.

In the entertainment industry, applications of voice similarity assessment include: voice casting [3], which is the reproduction of dialogue from a movie or video game, typically from one language to another; and voice assignment [4], which is the selection of a representative voice for a player's character, or avatar. Voice casting is currently a manual process, leaving a large scope for the introduction of automatic assistance. A recent study [3] proposed an automatic approach to voice casting by tagging speech samples with labels such as age, gender, voice quality and emotion, and applying a multi-label classification. The approach was shown to be effective at locating perceptually similar voices within a database. However, this approach is reliant on extensive manual labelling.

A challenge in automating voice similarity assessment is that the notion of 'similarity' is subjective. By using collective human judgments as a reference for what is typically perceived as similar, the effectiveness of various approaches can be established. Previous research on perceived voice similarity has pointed toward pitch, formant information and voice quality as being important cues [5, 6, 7]. In this study, we therefore consider the use of several automatically extracted phonetic features – 'auto-phonetic' features – for automatic voice similarity assessment using a current speaker recognition system. The following sections describe a pilot study involving automatic and listener experiments, followed by a discussion on the use of this proposed framework in practice.

2. Voice database

The 'Speakers In The Wild' (SITW) database [8], compiled for a special session at Interspeech 2016, was used as a source of speaker recordings. SITW was compiled for the study of speaker recognition in unconstrained conditions. As such, it contains speaker recordings in widely varied acoustic environments, microphone types, and speaking contexts. All recordings are in English, with a large variation in accent. From the full corpus, a subset of 175 speakers recorded on lapel microphones was selected for the subsequent experiments.

3. Automatic experiment

A range of auto-phonetic features were used in our experiments: F0, F1-F4, semitones of F0, along with first derivatives. F0 (Hz) is expected to model the range of F0 levels within which a speaker is phonating. To capture information relating to the intonation patterns used by a speaker, the associated derivative is included. F0 values converted to semitones compress the F0 range between-speaker and between-gender, improving intonation pattern modelling when used in combination with derivatives [9]. Long term formant (LTF) features have shown very promising results for forensic speaker recognition [10] along with their connection to perceived similarity [6]. Thus, F1-F4 and their derivatives were included.

The iVOCALISE automatic speaker recognition system [11] was used for speaker similarity experiments, with slight modifications to the features extracted in its 'auto-phonetic mode'. iVOCALISE operates within an i-vector PLDA (Probabilistic Linear Discriminant Analysis) framework. The modelling parameters include a UBM (Universal Background Model) of 256 components, with TV (Total Variability) and PLDA dimensions of 200 and 100 respectively.

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 using iVOCALISE. This was followed by a listener experiment to assess the relationship between comparison scores and perceived voice similarity.

4. Listener experiment

For each of the six target voices, a 'similar' cohort was formed by selecting the two closest voices in SITW, as ranked by the iVOCALISE comparison scores. Additionally, a 'different' cohort was formed by selecting two speakers at random, outside the top 10 highest scores, and a 'same' cohort was formed by selecting a different recording of the same speaker. A listener evaluation was created by pairing a sample of each target voice with a sample of each voice in its same, similar and different cohorts. Each sample was seven seconds in duration, and no samples were used more than once. With the six target voices, the test was therefore comprised of 30 comparisons, presented in a random order. An additional three comparisons were included to allow listeners to familiarise with the experiment. For each comparison, listeners were requested to judge the similarity of the two voices on scale of 1—9, ranging from 'very different' to 'very similar', while aiming to ignore the speaker accents, any non-speech noises, or any of the spoken content. The test was completed by 43 listeners (25 male, 18 female) via an online application, available at: http://oxfordwaveresearch.com/VocalSimilarityAssessmentv2.

5. Experimental Results

The full set of responses for male comparisons are shown in Figure 1. A significantly higher response value is observed for 'similar' voices automatically identified using iVOCALISE comparison scores. Additionally, there is a positive linear correlation (0.7, Pearson) between comparison score and median similarity rating for all male comparisons. In the female case however, the median response for similar comparisons is not significantly greater than that for different comparisons.

6. Discussion & Conclusion

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 significant promise. The performance of the proposed system was better with male voices than with female voices. Based on an evaluation by one of the authors, who is a trained phonetician, the highest-ranked female voices were judged to have very similar vocal timbre to their target voice, despite in several cases having different accents. Controlling for accents in a candidate dataset would reduce this bias. For voice parades, the degree of similarity must be fair, both to the suspect and the witness, whereas for media applications, the degree of similarity may be less important. With calibrated comparison scores, the degree of similarity could be inferred, allowing a suitable application-specific threshold to be set.

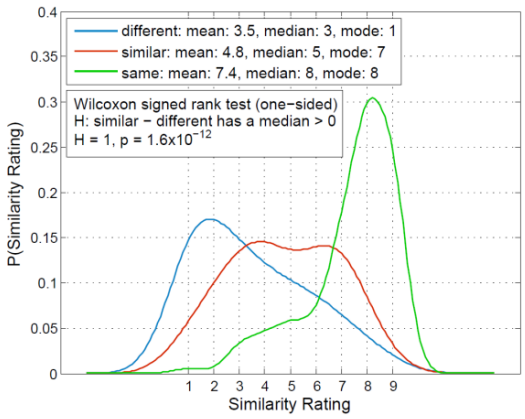

Figure 1: *Kernel density estimates of similarity ratings for male comparisons from all listeners (15*43 = 645 responses), where 1 indicates 'very different' and 9 'very similar'. The median response for 'similar' comparisons is significantly greater than the median for 'different' comparisons.*

With an appropriate database containing meta-data such as accent and age, or indeed additional voice quality labels, the approach of using an i-vector-based automatic system with phonetically-inspired features could prove an efficient and effective method of identifying potential cohorts of speakers for voice parades and voice casting.

References

[1] U.K. Home Office, "Advice on the use of voice identification parades", circular: 057/2003, December 2003.

[2] K. McDougall, "Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades", IJSLL, vol. 20, no. 2, pp. 163-172, 2013.

[3] N. Obin, A. Roebel and G. Bachman, "On automatic voice casting for expressive speech: Speaker recognition vs. speech classification", In proceedings of ICASSP 2014, Florence, Italy, pp. 950-954, 2014

[4] Y. Adachi, S. Kawamoto, T. Yotsukura, S. Morishima and S. Nakamura, "Automatic voice assignment tool for Instant Casting movie System," In proceedings of ICASSP 2009, Taipei, pp. 1897-1900, 2009.

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

[6] E. Zetterholm, M. Blomberg and D. Elenius, "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, 2010.

[7] J. Lindh and A. Eriksson, "Voice similarity — a comparison between judgements by human listeners and automatic voice comparison", In proceedings of FONETIK 2010, pp 63-69, 2010.

[8] M. McLaren, F. Luciana, D. Castan and A. Lawson, "The Speakers in the Wild (SITW) Speaker Recognition Database", submitted to Interspeech 2016, San Francisco, U.S.A., 2016.

[9] D. R. Ladd, "Intonational Phonology", Cambridge University Press, 1996.

[10] W. Heeren, D. van der Vloed and J. Vermeulen, "Exploring long-term formants in bilingual speakers". In Proceedings of IAFPA 2014, Zürich, Switzerland, 2014.

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