Speaker recognition with Phonetic and Automatic Features using VOCALISE software

Anil Alexander[1], *Oscar Forth*[1], *Marianne Jessen*[2] *and Michael Jessen*[3]
[1]*Oxford Wave Research Ltd, Oxford, United Kingdom*
{anil|oscar}@oxfordwaveresearch.com
[2]*Stimmenvergleich, Wiesbaden, Germany*
jessen@stimmenvergleich.de
[3]*Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany.*
michael.jessen@bka.bund.de

In this article, we present an innovative new system for forensic speaker recognition that provides the capability to perform comparisons using both 'traditional' forensic phonetic parameters and 'automatic' spectral features in a semi- or fully automatic way. This speaker recognition system called VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence) allows the forensic practitioner to statistically model and compare long-term formant information (Nolan & Grigoras 2005) and formant dynamics (McDougall 2005), along with spectral features like Mel Frequency Cepstral Coefficients (MFCCs). It is capable of comparing phonetic and automatic features from a test audio file from a target speaker against features from an audio file of a suspected speaker or an entire list of suspected speakers, and produces a likelihood score for each comparison. VOCALISE seeks to form a bridge between traditional forensic phonetics-based speaker recognition and forensic automatic speaker recognition and provides a coherent means of expressing the combined results.

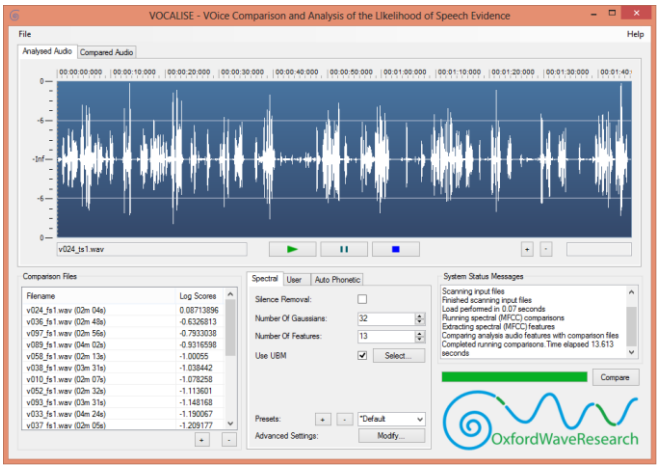

Figure 1. VOCALISE performing a one-to-many speaker comparison in spectral mode

Particular attention was given to its capability of providing a common methodological platform for both classical automatic and phonetic speaker recognition. Three operation modes called 'spectral', 'user', and 'auto phonetic' are currently included in VOCALISE (Figure 1). Spectral refers to the automatic extraction of the kind of features that are most commonly used in automatic speaker and speech recognition (currently MFCCs). User (-defined) refers to the option that lets the users use their own stream(s) of values which can be manually measured, labelled, or corrected such as formant frequencies, fundamental frequency, or durations of

sounds, syllables or sub-syllabic constituents (units relevant to tempo and rhythm), or even auditory features. Auto-phonetic refers to the automatic (unsupervised) extraction of phonetic features (currently formants F1 to F4 selected in any combination for analysis). VOCALISE allows for normalisation and extraction of dynamic information, Gaussian Mixture Modeling (GMM), as well as the creation of statistical models for populations using universal background models (UBMs) for phonetic, spectral or user-defined features interchangeably.

This system has been tested against good quality databases like Pool 2010 (Jessen et al. 2005) and DyVIS (Nolan et al. 2009) as well as telephone-quality real casework data. VOCALISE has been tested against real case data in analysing the speech of 22 male speakers of German from authentic anonymised cases. As would be expected, in real-case data, there were many detailed behavioural and technical differences within and between files, and several of the files had signal distortions and other quality problems. The language used was not homogeneous comprising of different regional and ethnic varieties of German but without strong dialects or heavy foreign accents. The initial experiments with subsets of good quality databases like Pool 2010 (22 speakers) and DyVIS (17 speakers) have been very promising, achieving close to complete speaker separation (0.1% and 0.374% equal error rate (EER) respectively, using spectral methods). The auto-phonetic mode yielded approximately 8.9% EER for Pool 2010 and 5.88% for DyVIS. Using the real case files containing between 20 and 60s of speech, initial testing obtains an EER of 12.6% in the spectral mode. The settings for this result included 32 Gaussians components, 13 MFCCs, Mean Variance Normalisation (MVN), Symmetric Testing (inverting model speaker and test speaker in calculating similarity scores). Long-term formant analysis of the same data was performed using the auto-phonetic module of VOCALISE. Using 6 Gaussians, MVN, Symmetric Testing and delta features (derivatives of formant values providing formant dynamic information), EER was at 18.1%. Considering that this is about fully natural, quality reduced speech, unsupervised formant tracking, and no other information than F1 to F3, this is a promising result.

This system provides capabilities that not only makes it possible to apply classical automatic speaker recognition, but also to analyse the speaker-discriminative information of acoustic phonetic data such as formant frequencies, fundamental frequency or sound durations. Whereas features pertaining to the spectral envelope such as MFCCs are powerful, they are also very sensitive to channel effects and recording quality, and are mostly data-driven and less directly connected to the theory of speech production (Rose 2002). Processing phonetic data will be in many ways complementary and will offer insights into the voice comparison analysis that the classical automatic methods cannot.

References

Jessen, M., O. Köster & S. Gfroerer (2005). Influence of vocal effort on average and variability of fundamental frequency. *Intern. J. of Speech, Language and the Law,* **12**, 174–213.

McDougall, K. (2005). *The Role of formant dynamics in determining speaker identity*. Ph.D. Dissertation, University of Cambridge.

Nolan, F. & C. Grigoras (2005). A case for formant analysis in forensic speaker identification. *Intern. J. of Speech, Language and the Law,* **12**, 143–173.

Nolan, F., K. McDougall, G. de Jong & T. Hudson (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Intern. J. of Speech, Language and the Law* **16**: 31–57.

Rose, P. (2002). *Forensic speaker identification*. London: Taylor & Francis.