VOCALISE: A forensic automatic speaker recognition system supporting spectral, phonetic, and user-provided features

Anil Alexander, Oscar Forth, Alankar Aryal Atreya and Finnian Kelly

Research and Development
Oxford Wave Research Ltd, United Kingdom
{anil¦oscar¦alankar¦finnian}@oxfordwaveresearch.com

Abstract

In this article we present the latest version of VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence), a forensic automatic system for speaker recognition. VOCALISE, with selectable state-of-the-art and legacy speaker modelling algorithms allows the forensic practitioner to work with spectral features (such as Mel Frequency Cepstral Coefficients (MFCCs)), phonetic features (such as formants), or features of their own choice (such as voice quality metrics, articulation rate, etc.). It is capable of comparing features from a test audio file of 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 or likelihood ratio for each comparison. It is built with an 'open-box' architecture that transparently allows the user to provide their own data to train the system's algorithms. These algorithms include Gaussian Mixture Modelling (GMM) with (and without) MAP (maximum a posteriori) adaptation, i-vector extraction with PLDA (Probabilistic Linear Discriminant Analysis) and cosine distance comparison. VOCALISE seeks to form a bridge between traditional forensic phonetics-based speaker recognition and forensic automatic speaker recognition.

1. Introduction

VOCALISE is a forensic automatic speaker recognition system that allows users to perform comparisons using both 'traditional' forensic phonetic parameters and 'automatic' spectral features in a semi- or fully automatic way.

1.1. Motivation

Currently, the majority of forensic speaker recognition case-work across the world is performed by forensic phoneticians. These practitioners often possess a wealth of knowledge and experience of phonetic and linguistic voice analysis and a deep understanding of the legal requirements in their countries. They find themselves 'out of the loop', or unable to leverage their know-how in a fully automatic analysis. Some of these experts would like to make their analysis more objective using likelihood ratios but find it difficult to estimate the relevant statistics of the potential population and suspected speakers for the features they are analysing. We sought to develop a software system which would allow the practitioner to perform both fully automatic analysis based on spectral features as well as expert-aided analysis based on phonetic features with the capability of providing similarity scores or likelihood ratios.

1.2. History

The development of VOCALISE, originally based on Gaussian mixture modelling (GMM) and Mel-frequency cepstral coefficients, began in 2012 [1]. Development has been ongoing since, with the inclusion of capabilities such as long-term distributions of automatically extracted phonetic features, including formants and user-provided features, as well as selective processing of annotated regions [2]. More recently, in 2015, a version incorporating i-vector modelling [3] called iVOCALISE has been developed. Some of this development has benefitted from the support and collaboration of the German Bundeskriminalamt (BKA), the Netherlands Forensic Institute (NFI), as well as the UK Ministry of Defence.

1.3. Design Philosophy

One of the major criticisms of the application of automatic speaker recognition is that the underlying algorithms form a black box into which the forensic examiner is unable to look, or indeed adapt to their own requirements. VOCALISE has been developed with an 'open-box' architecture. The idea underlying the design is that the user should be able to change the system parameters and introduce new data at every step of the speaker recognition process. With this approach, the user is not limited to manufacturer-provided models or configurations, and has the ability to train the system specifically for their problem domain.

We have attempted to open the black box for the user by allowing flexibility in the choice of features and the parameters for feature extraction, the modelling

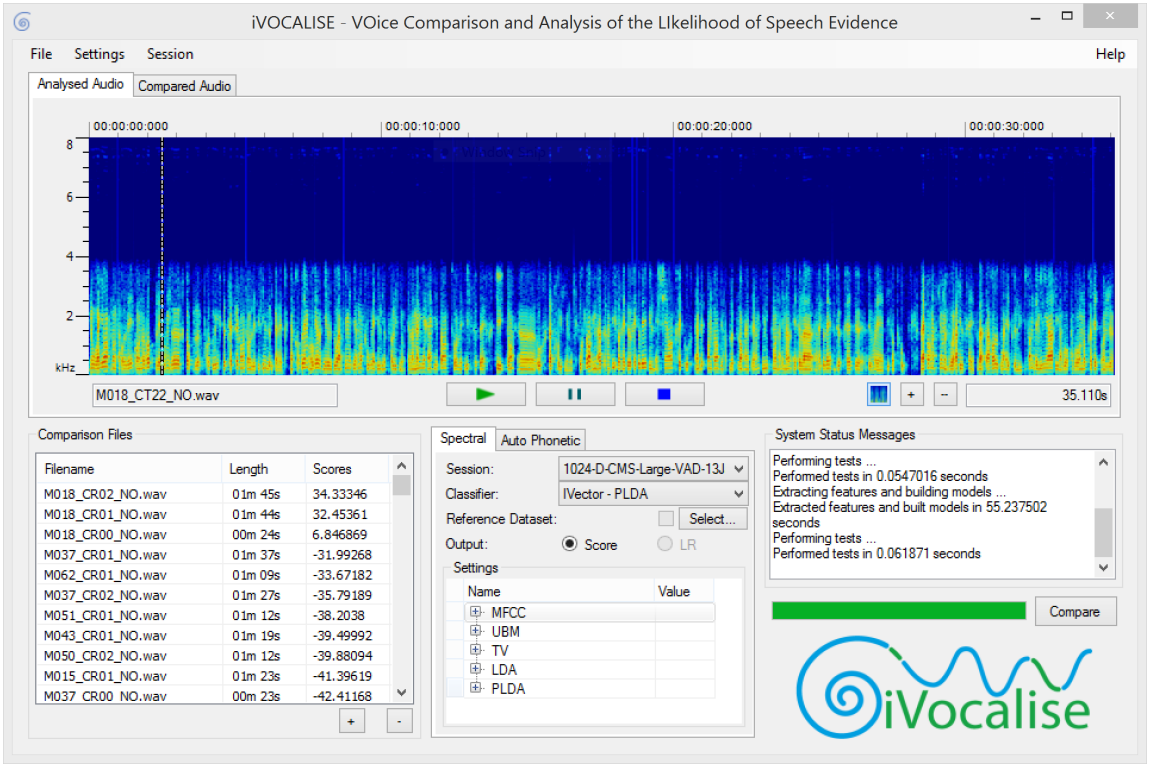

Figure 1: VOCALISE main user interface: Comparison performed using i-vector PLDA and spectral (MFCC) mode

techniques and in the calculation of results. In addition, VOCALISE interfaces with 'trusted' programs like Praat [4] to allow the user to utilise features provided by such specialist software.

2. Contents of the 'open' box

VOCALISE is supplied with a default configuration consisting of pre-trained models that have been tested and optimised. Should the user wish to customise the configuration, there is flexibility at all stages of the speaker recognition process for both the state of the art as well as legacy algorithms.

2.1. Feature extraction options

Information important for speaker discrimination is first extracted from the speech signal by conversion into a set of features. These features can either be automatically or manually extracted as described below.

2.1.1. Spectral Features

Spectral features are descriptors of the frequency characteristics which are automatically extracted from a speech sample over short time windows, and are the most commonly used feature-type in speech and speaker recognition. VOCALISE currently supports flexible MFCC features with:

- Adjustable frequency band selection.
- Optional energy, delta and delta-delta coefficients.
- Cepstral mean subtraction (CMS) and variance normalization (CMVN).

2.1.2. Auto phonetic Features

The use of 'auto-phonetic' features, i.e. phonetic features extracted in an automatic (unsupervised) way, is supported via an interface with Praat [4]. Currently, any combination of formants F1 to F4 can be selected via the user interface. Other auto-phonetic features, such as pitch, can be included by modifying an external Praat script.

2.1.3. User Provided Features

'User-provided' refers to the option that allows users to provide their own features to the system [2]. These may be features that have been manually measured and labelled, such as a hand-corrected formant tracks, or other features such as voice quality metrics, articulation rates, durations of sounds, syllables or sub-syllabic constituents (units relevant to tempo and rhythm), or even auditory features. Such features can be provided as input files as columns of data in text format.

2.2. Modelling algorithms provided

The system includes three main feature modelling algorithms that can be applied to both spectral as well as auto-phonetic features. These speaker identity training algorithms include Gaussian Mixture Modelling (GMM) with (and without) MAP (Maximum a posteriori) adaptation, i-vector extraction with PLDA (Probabilistic

Linear Discriminant Analysis) and cosine distance comparison. The user can switch between the various modelling techniques in VOCALISE easily.

2.3. Ability to measure performance metrics using large many to many comparisons

The recent ENFSI Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition [5] recommend that practitioners quantify the speaker discriminatory performance of recordings that are representative of their casework. As it is important for the expert to measure performance, VOCALISE allows the user to conduct large-scale many to many (NxM) speaker comparisons and quickly obtain various performance values and representations. One such representation is shown in the figure below.

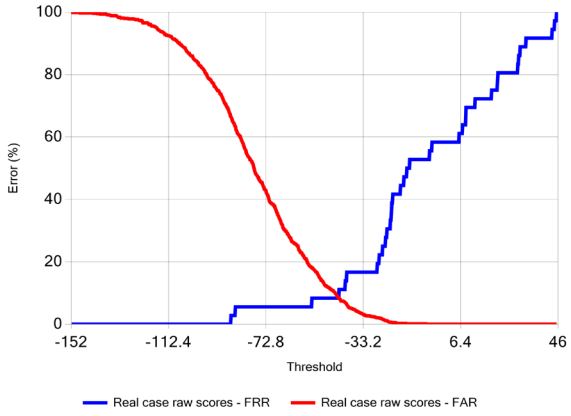

Figure 2: An example equal error graph with raw i-vector comparison scores obtained using the German real case dataset discussed in Section 4.

3. iVOCALISE:- the i-vector framework

The latest version of VOCALISE (called iVOCALISE) operates using an i-vector PLDA (Probabilistic Linear Discriminant Analysis) framework [3, 6], which has emerged as the dominant approach in high-performing speaker recognition systems.

This approach offers performance improvements over its predecessors, including GMM-UBM, particularly where there is a significant acoustic mismatch between the samples under comparison (different recording channels or different languages, for example). In this framework, a sample of speech is converted into feature vectors (Section 2.1), and subsequently a low-dimensional, fixed-length representation known as an i-vector. The conversion from speech sample to the final i-vector attempts to preserve speaker-specific information, discarding as much as possible, information not related to the identity of the speaker. In iVOCALISE, i-vectors obtained from two speech samples can be compared using a cosine distance measure, or PLDA [6], which computes the likelihood that the pair of i-vectors originate from the same speaker versus different speakers. In keeping with the open-box philosophy, iVOCALISE allows the user to customise the i-vector framework by introducing their own data at multiple stages of the system, and by tuning the modelling parameters of the UBM (Universal Background Model), the TV (Total Variability) model, along with LDA (Linear Discriminant Analysis) and PLDA.

3.1. Training the system

VOCALISE provides ready-to-use `sessions', consisting of pre-trained and optimised models. If desired, users can create their own custom session from scratch or from an existing session, by introducing data and setting modelling parameters for their particular use-case.

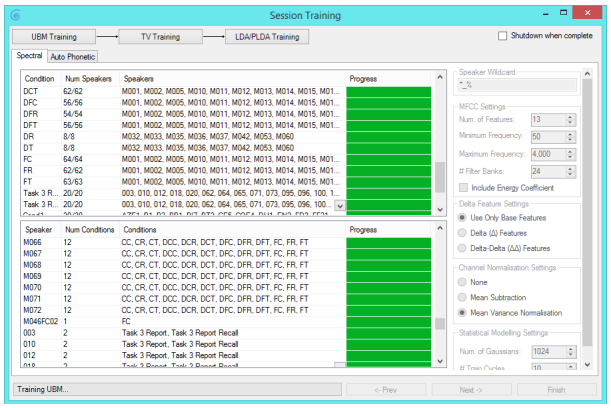

Figure 3: VOCALISE session creation: Training the i-vector system with the user's own datasets

Users should be aware however, that the ability of the i-vector PLDA framework to perform reliably is dependent on the use of appropriate development data. Here we provide some general guidelines for custom training of the VOCALISE i-vector system.

3.1.1. Training the UBM

The UBM (Universal Background Model) is a representation of a global `acoustic space'. Training recommendations:

- Data Quantity: A large total duration of speech is generally advised: approximately 10 hours or more, after removal of silences/non-speech segments, is recommended.
- Data Quality: Data diversity is important. UBM training data should originate from a large number of speakers; we recommend 200 or more. Increasing the diversity of the UBM data by including speakers of different genders, ages and languages, and by including recordings in different environments, is beneficial to training a system that performs well across a range of use-cases.

3.1.2. *Training the TV model*

The TV (Total Variability) model is the final stage in converting a speech sample to an i-vector. We apply the same recommendations to TV training as to UBM training, with a particular emphasis on ensuring data diversity. By default, UBM data is also used for TV training.

3.1.3. *PLDA model training*

PLDA models within-speaker and between-speaker variability, which is then used to calculate a likelihood score from two i-vectors under comparison. LDA (Linear Discriminant Analysis) is applied to i-vectors prior to PLDA, to reduce dimensionality and enhance separability. Again, the UBM training guidelines apply, with the following additional comments:

- Data must have speaker labels (not required for UBM or TV training).
- At least two recordings per-speaker are required, but it is recommended to include as many as possible.
- It is beneficial to include speech samples from speakers and recording conditions that are relevant to the desired use-case.

4. Initial experiments with real case data

The iVOCALISE system has been used with real forensic data, including the NFI-FRITS corpus that contains real telephone intercept data, as described in [7], and the German real-case corpus from the BKA, referred to in [8]. While it is possible to train and adapt the system to context-specific data, simply using a pre-trained session file (containing UBM, TV matrix and PLDA information) that was trained with spectral MFCC data, (using delta parameters, 1024 Gaussians, and a 400 element i-vector), VOCALISE has obtained promising results when compared to those previously reported. It is possible for the user to further adapt the system to their specific contexts if required.

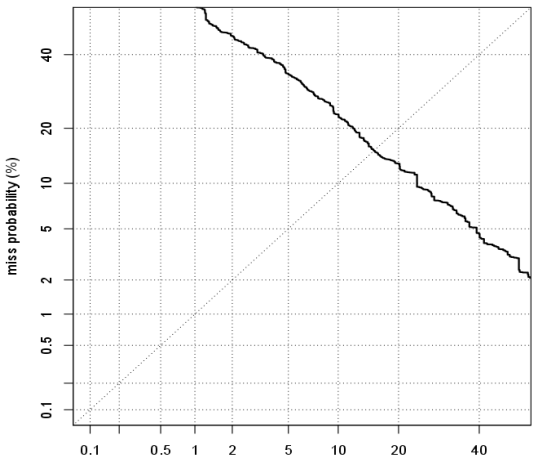

Figure 4: DET Curve obtained using the NFI-FRITS dataset using a spectral session file

With the NFI-FRITS dataset, the DET curve obtained from all language comparison is shown in Figure 4. The corresponding overall equal error rate (EER) is comparable to that presented in [7]. In the German real-case data [8], the iVOCALISE system obtained an EER of 6.89% using a pre-trained session file. This result is betters than the best performing system with this dataset as described in [8].

5. Conclusions

VOCALISE provides the forensic practitioner with an automatic speaker recognition system capable of performing comparisons using both 'traditional' forensic phonetic parameters and 'automatic' spectral features in a semi- or fully automatic way. It enables the user to make objective estimates of the strength of the evidence in a speaker recognition case. Processing phonetic data will be in many ways complementary and will offer insights into voice comparison analysis that classical automatic methods cannot. We seek to provide a unified 'open-box' architecture, and return control into the hands of the expert users.

6. References

[1] M. Jessen, O. Forth and A. Alexander, "VOCALISE: Eine gemeinsame Plattform für die Anwendung automatischer und semiautomatischer Methoden in forensischen Stimmenvergleichen" *Polizei & Wissenschaft* vol. 4/2013, pp. 2-19, 2013.

[2] M. Jessen, A. Alexander and O. Forth, "Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software.", Proceedings of the Audio Engineering Society 54th International Conference; London, pp. 28–35, 2014.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Transactions on Audio, Speech & Language Processing, vol. 19, no. 4, pp. 788–798, 2011.

[4] P. Boersma and D. Weenink :"Praat: doing phonetics by computer [Computer program]" Version 5.3.42, retrieved 2 June 2013 from http://www.praat.org/.

[5] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen and T. Niemi, "Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition", Frankfurt: Verlag für Polizeiwissenschaft, 2015.

[6] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in IEEE 11th International Conference on Computer Vision (ICCV), pp. 1–8, 2007.

[7] D. van der Vloed, J., Bouten and D. van Leeuwen, "NFI-FRITS: A forensic speaker recognition database and some first experiments. Proc. Odyssey 2014, 6-13, 2014.

[8] Y. T. Solewicz, G. J. Becker and S. Gfroerer, "Comparison of speaker recognition systems on a real forensic benchmark.", Proceedings of Odyssey 2012 Singapore, 2012.