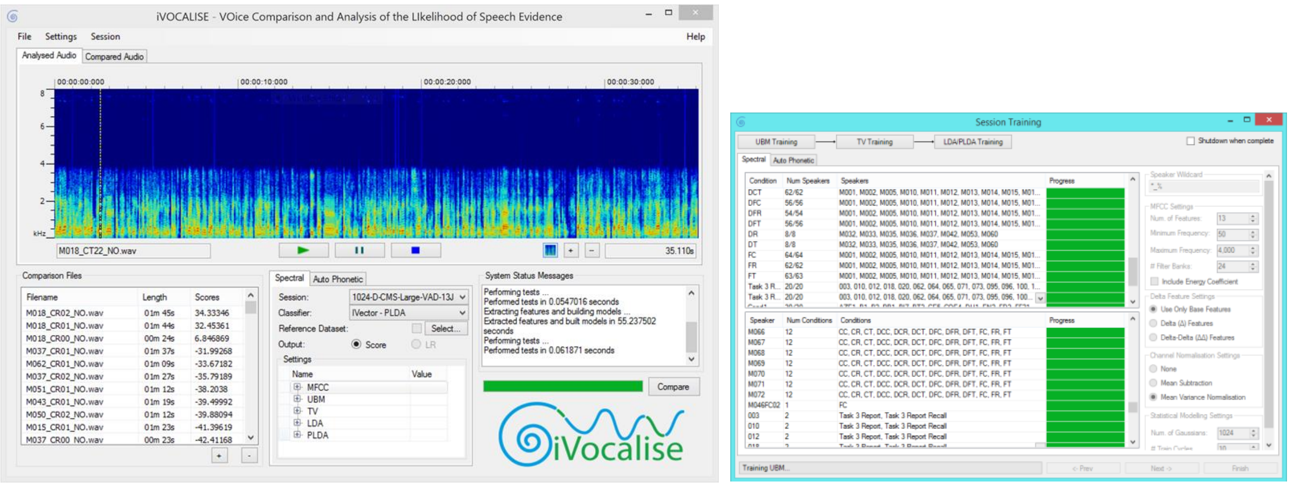
iVOCALISE: An i-vector-based automatic speaker recognition system using spectral and phonetic features

Anil Alexander, Oscar Forth, Alankar A. Atreya and Finnian Kelly
Research and Development, Oxford Wave Research Ltd, Oxford, United Kingdom
{anil|oscar|alankar|finnian}@oxfordwaveresearch.com

In this article we present a novel framework based on ‘i-vectors’ for extracting phonetic or spectral features of the voice and representing them as individual points in a compact lower dimensional space. This represents a generational change from the last version of VOCALISE, an automatic speaker recognition system, originally based on Gaussian mixture modelling (GMM) and Mel-frequency cepstral coefficients. Within the GMM-based system we have previously developed capabilities such as modeling long-term distributions of automatically extracted phonetic features, including formants and user-provided features, as well as selective processing of annotated regions.

In the latest version of VOCALISE (called iVOCALISE) we have incorporated the dominant approach in high-performing algorithms, namely an i-vector PLDA (Probabilistic Linear Discriminant Analysis) framework (described in Dehak et al, 2011), and have applied this to both phonetic and spectral features. The i-vector-based approach represents a significant performance improvement over GMMs, particularly when there is significant mismatch in the recording conditions of the samples under comparison. The conversion from speech sample to i-vector attempts to preserve speaker-specific information and discard information not related to the identity of the speaker. The i-vector approach has been extensively explored in literature, particularly with regard to spectral features such as MFCCs. Phonetic features such as formants are not commonly analysed within this framework however. We have sought to provide exactly the same i-vector-based framework to phonetic features.



Comparison Files

Filename	Length	Scores
M018_CR02_NO.wav	01m 45s	34.33346
M018_CR01_NO.wav	01m 44s	32.45361
M018_CR00_NO.wav	00m 24s	6.846869
M037_CR01_NO.wav	01m 37s	-31.99268
M062_CR01_NO.wav	01m 09s	-33.67182
M037_CR02_NO.wav	01m 27s	-35.79189
M051_CR01_NO.wav	01m 12s	-38.2038
M043_CR01_NO.wav	01m 19s	-39.49992
M050_CR02_NO.wav	01m 12s	-39.88094
M015_CR01_NO.wav	01m 23s	-41.39619
M037_CR00_NO.wav	00m 23s	-42.41168

Session: 1024-D-CMS-Large-VAD-13J
Classifier: IVector - PLDA
Reference Dataset: □
Output: ● Score ○ LR

Settings

Name	Value
⊞ MFCC	
⊞ UBM	
⊞ TV	
⊞ LDA	
⊞ PLDA	

System Status Messages

```
Performing tests ...
Performed tests in 0.0547016 seconds
Extracting features and building models ...
Extracted features and built models in 55.237502
seconds
Performing tests ...
Performed tests in 0.061871 seconds
```

□ Shutdown when complete

Condition	Num Speakers	Speakers	Progress
DCT	62/62	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
DFC	56/56	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
DFR	54/54	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
DFT	56/56	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
DR	8/8	M032, M033, M035, M036, M037, M042, M053, M060	🟩
DT	8/8	M032, M033, M035, M036, M037, M042, M053, M060	🟩
FC	64/64	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
FR	62/62	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
FT	63/63	M001, M002, M005, M010, M011, M012, M013, M014, M015, M01...	🟩
Task 3 R...	20/20	003, 010, 012, 018, 020, 062, 064, 065, 071, 073, 095, 096, 100, 1...	🟩
Task 3 R...	20/20	003, 010, 012, 018, 020, 062, 064, 065, 071, 073, 095, 096, 100...	🟩

Speaker	Num Conditions	Conditions	Progress
M066	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M067	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M068	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M069	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M070	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M071	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M072	12	CC, CR, CT, DCC, DCR, DCT, DFC, DFR, DFT, FC, FR, FT	🟩
M046FC02	1	FC	🟩
003	2	Task 3 Report, Task 3 Report Recall	🟩
010	2	Task 3 Report, Task 3 Report Recall	🟩
012	2	Task 3 Report, Task 3 Report Recall	🟩

Speaker Wildcard: *_%

MFCC Settings
Num. of Features: 13
Minimum Frequency: 50
Maximum Frequency: 4,000
Filter Banks: 24
□ Include Energy Coefficient

Delta Feature Settings
● Use Only Base Features
○ Delta (Δ) Features
○ Delta-Delta (ΔΔ) Features

Channel Normalisation Settings
○ None
○ Mean Subtraction
● Mean Variance Normalisation

Statistical Modelling Settings
Num. of Gaussians: 1024

Training UBM...

Figure 1. The iVOCALISE spectral comparison user interface (left), with user-provided data being used to train the universal background model, total variability matrix, and linear discriminant analysis (right).

This system has been developed with an ‘open-box’ architecture in which the user may adjust most of the feature and modelling parameters, and introduce new data at every step of the speaker recognition process. The user is, by design, not limited to manufacturer-provided models or configurations. They can either train the i-vector system and corresponding background models from scratch, or adapt the provided models with their own data.

The iVOCALISE system has been used with real forensic data, including the NFI-FRITS corpus that contains real telephone intercept data (van der Vloed et al, 2014), and the German real-case corpus from the BKA (Solewicz et al, 2012). 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, (with delta coefficients, 1024 Gaussians, and a 400 element i-vector), iVOCALISE has obtained promising results when compared to those previously reported. In the German real-case data, the iVOCALISE system obtained an equal error rate (EER) of 6.89% using a pre-trained session file. This result is better than the highest-performing system with this dataset as described in Solewicz et al, 2012. iVOCALISE demonstrates a successful application of the i-vector framework to both phonetic and spectral features.

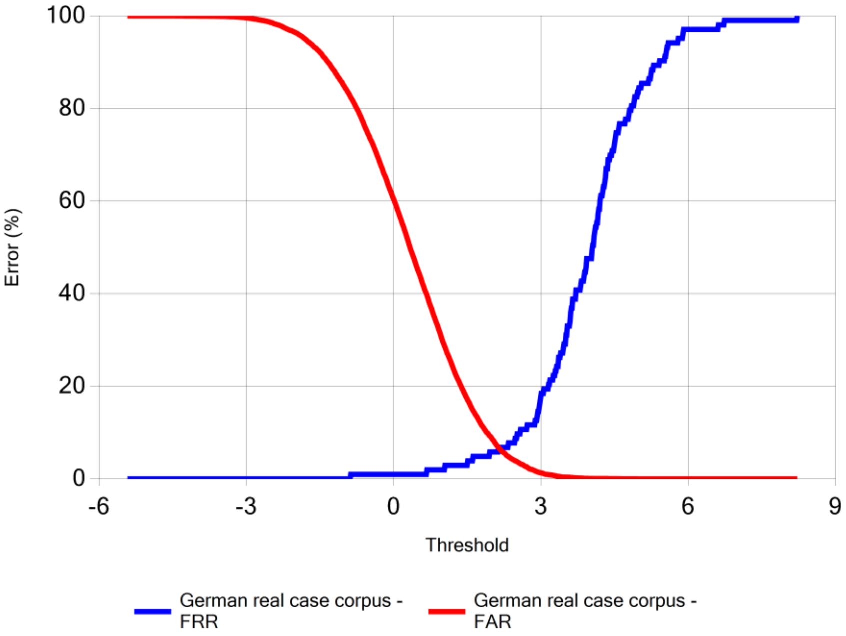


Figure 2. An example equal error graph with raw i-vector comparison scores obtained using the German real case dataset

References

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

Jessen, M. A., Alexander, A. and Forth, O. (2014). Forensic voice comparisons in German with phonetic and automatic features using VOCALISE software*. In proceedings of the Audio Engineering Society 54th International Conference*, London, pp. 28–35.

Solewicz, Y. T., Becker, G. J. and Gfroerer, S. (2012). Comparison of speaker recognition systems on a real forensic benchmark. *In proceedings of Odyssey 2012*, Singapore.

van der Vloed, D., Bouten, J. and van Leeuwen, D. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. *In Proceedings of Odyssey 2014*, Joensuu, Finland, pp. 6-13.