Audio Engineering Society

Conference Paper 4

Presented at the AES 8th International Conference on Audio Forensics
2024 June 27-29, Denver, Colorado, USA

Handling real-world challenges of variable speech quality and multiple speakers in forensic automatic speaker recognition using VOCALISE

Anil Alexander[1], Linda Gerlach[1], Thomas Coy[1], Oscar Forth[1], and Finnian Kelly[1]

[1]Oxford Wave Research Ltd, Oxford, United Kingdom

Correspondence should be addressed to Anil Alexander (anil@oxfordwaveresearch.com)

ABSTRACT

Recordings presented for forensic speaker recognition comparison generally have a variety of complexities that must be taken into account before analysis. These could involve the quality of the recording, the duration and linguistic diversity in the speech of the speaker of interest, and the presence of other speakers in the recording. One of the first tasks that the forensic analyst has to perform is to separate out a sample of speech from only the speaker of interest, and make sure there is an adequate quantity and quality to create a voice model for an automatic comparison. This task is not straightforward and requires experience and skill from the practitioner to create a reliable model. Traditionally, automatic speaker comparison implicitly assumes that input data contains only one speaker per recording or has been preprocessed to contain only one speaker. However, depending on the nature of the case, there may be a varying number of speakers in different recordings and forensic practitioners may not have the capacity to preprocess many files, each potentially containing multiple unknown speakers. The quality of a recording effectively depends on both the acoustic signal quality as well as the quantity and diversity of linguistic content present in the recording. In determining whether a file can be used for forensic analysis, the practitioner has to frequently rely on their often-subjective determination of the quality of the recording. For instance, they may decide that a recording is very noisy and contains very little speech, and therefore unlikely to be useful for analysis. It would be very helpful to provide the practitioner with some objective quality metric information so that they can decide whether to proceed with their analysis, and to indicate what the potential error rates might be. In this article, using the latest version of VOCALISE (Voice Comparison and Analysis of the Likelihood of Speech Evidence), a forensic automatic speaker recognition system, we present some pragmatic solutions to some of these common case-related issues faced by forensic practitioners. For the objective analysis of the acoustic quality of different recordings we consider a 'star' rating audio quality metric (from 1 to 5) that includes net-speech duration, signal-to-noise ratio, and amount of clipping in the signal. We also compare two options for handling multi-speaker files including using manual selections from a multi-speaker file and an automatic segmental mode, which splits recordings into segments of an adjustable length and overlap. Segmental mode facilitates the task of determining whether a multi-speaker recording contains the speech of a specific speaker, and at what point in the recording their speech occurs, in a fully automatic way. Using the forensically relevant WYRED speaker recognition database, we demonstrate the effect of comparing files of different 'star' quality ratings and examine the error rates obtained by using three different approaches to handling multi-speaker recordings.

1 Introduction

Forensic practitioners employing automatic speaker recognition techniques are regularly faced with the challenges of data selection, management, and processing. In forensic and investigative speaker recognition it is not unusual to have a large number of files to analyse. These recordings may also come in a variety of different formats and recording conditions and will need to be manually preprocessed before speaker recognition analysis. Preprocessing could include the conversion of file types, extracting the audio component from video files, and selecting segments of speech for the speaker of interest. It may be necessary to analyse large volumes of multi-speaker recordings in different recording formats and conditions to identify which files may contain a speaker of interest. In time-critical investigations, if manual processing is required it could slow down the whole analysis. Automatic file analysis and processing methods could help speed up this preliminary investigative analysis.

The VOCALISE forensic automatic speaker recognition system, which is widely used for forensic and investigative applications, enables a forensic practitioner to perform speaker comparisons in a flexible way, and to estimate likelihood ratios under the same-speaker and different-speaker hypotheses in order to evaluate the strength of the evidence in a speaker comparison case. VOCALISE is built on an 'open-box' philosophy, allowing the practitioner to look into the various feature extraction and speaker modelling algorithms and to transparently observe and visualise voice models for different speakers. The system's core algorithm is an 'x-vector' based PLDA framework that uses Deep Neural Networks (DNNs) to obtain compact representations of speaker voices. This DNN-based approach has significantly improved recognition accuracy in adverse conditions typical to forensic and law enforcement cases compared to previously used 'i-vectors' [1].

In this paper we consider two common issues affecting the workflow of forensic or investigative speaker recognition analysis, namely objectively quantifying the quality of a recording prior to analysis, and efficiently handling multi-speaker recordings in time-critical analysis. Various approaches to address these issues using functionality available within the VOCALISE system are considered, and the effect on accuracy is examined using a forensically relevant speaker recognition database.

2 Audio Quality Profiling

When performing automatic (or human-driven) speaker recognition, the forensic practitioner is more likely to obtain better results when using good quality data. Using poor-quality recordings in speaker comparisons can lead to misleading outcomes as the voice of a speaker of interest may be significantly masked or degraded, resulting in a poor-quality voice model for the speaker. The impact of quality variation has previously been documented in the automatic speaker recognition research domain, for example [2, 3].

Audio quality profiling provides forensic practitioners with valuable insights into their data using reliable, numerical metrics. Making reference to such metrics can inform the practitioner's decision-making process when it comes to obtaining the best voice model for a speaker. It further enables forensic practitioners to capture the technical conditions in a case dataset and then easily select recordings most representative of these conditions from a broader speaker pool to form a relevant population or validation test set. Furthermore, quality metrics can contribute to report writing by providing repeatable, scientific grounds for data selection and corroborating or explaining observations from auditory analyses.

Within VOCALISE, each recording is given a star rating, ranging from 1 to 5 stars, as a holistic measure of the recording's suitability for automatic comparison. The star rating is informed by three metrics: net-speech duration (total speech duration post-voice activity detection, when pauses and silences are removed), WADA SNR (waveform amplitude distribution analysis signal-to-noise ratio) [4], and clipping. Clipping is measured as the percentage of 100 ms frames that contain at least 25% of samples at the minimum or maximum value of the samples in the frame. Table 1 summarises the default star rating heuristic used in VOCALISE.

Metric	★☆☆☆☆	★★☆☆☆	★★★☆☆	★★★★☆	★★★★★
Net Speech (s)	> 0.00	> 10.00	> 15.00	> 20.00	> 30.00
WADA SNR (dB)	> -100.00	> 6.00	> 12.00	> 18.00	> 24.00
Clipping (%)	< 100.00	< 50.00	< 50.00	< 50.00	< 50.00

Table 1. Star rating metrics and default thresholds.

The rating is determined by the lowest value across the three metrics, on the basis that the quality of the file can only be as good as its worst attribute. For example, if a recording is of less than 10 seconds of net speech and has a WADA SNR of more than 24 dB

and 0% Clipping, the star rating will be 1. The default star rating thresholds have been informed by initial explorations of audio quality [5].

The Minimum Audio Quality threshold for files, i.e. the minimum star rating required to be submitted for comparison, may be configured by the user according to their specific use case requirements. Furthermore, thresholds can be adjusted for each of the metrics separately by adjusting the minimum net-speech duration, minimum WADA SNR, and minimum clipping percentage. As such, only files meeting these criteria will be used for comparison, while others will be automatically excluded by VOCALISE. Sometimes it may be necessary to use lower-quality recordings. In such cases, it is important to contextualise the results based on objective audio quality metrics.

3 Multi-Speaker Recordings

In forensic casework it is common to encounter recordings containing speech from multiple speakers. In order to expedite preliminary preprocessing in time-critical analysis, we consider three different approaches to efficiently process multi-speaker recordings. These approaches are discussed below.

3.1 Naïve Comparison

'Naïve comparison' is a minimal effort approach that involves disregarding the presence of multiple speakers in a recording and proceeding with a one-to-one, one-to-many, or many-to-many speaker comparison. This approach may be effective in cases where the other speakers only contribute few and short utterances or are not distinctly audible and can be considered background noise. Naïve comparison, although not as effective as more selective approaches, can, rather surprisingly, also yield reasonable speaker discrimination even when there is a balanced split of speech from two speakers in the recording.

Figure 1 shows the VOCALISE interface for a one-to-many naïve comparison for recordings containing only one speaker each. Spectrograms or waveforms of the loaded recordings can be displayed, and audio played back. Comparison files are ranked based on the scores against the analysis file, as shown in the lower left box.

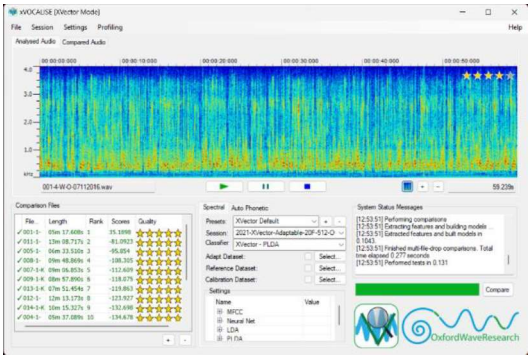

Figure 1. VOCALISE interface with results of a one-to-many comparison.

3.2 Selection of Regions

The most common method for dealing with multi-speaker files is for the forensic practitioner to manually preprocess the file and submit only files containing one speaker for speaker comparisons. This is normally done manually using audio editing software. Whereas in previous versions of VOCALISE it was necessary that users provided files that contained only a single speaker, practitioners can now perform manual selections of a multi-speaker file within the software to isolate a specific speaker in the recording.

After loading or converting data in the audio management panel, the user can then open an audio selection dialog for the file in question. The user is presented with a window in which audio can be played back and viewed as a zoomable spectrogram or waveform. Here, subsections of the audio can be selected. The selections are demarcated by clicking and dragging the cursor across a region of the spectrogram or waveform where the speech of the target speaker exists, before adding to an aggregated list of selections across the file. Completing this process will create a 'selections file', which lists the start and end timestamps for the selected regions. When files with selections are used in comparisons, only the concatenated selections are evaluated, thus isolating speech from a single speaker. Though presented here as a multi-speaker solution, this same technique is also effective when seeking regions of higher quality speech in a recording containing only a single speaker.

Figure 2 displays the VOCALISE selections window. The highlighted regions within the spectrogram mark the selections that will be considered in a comparison.

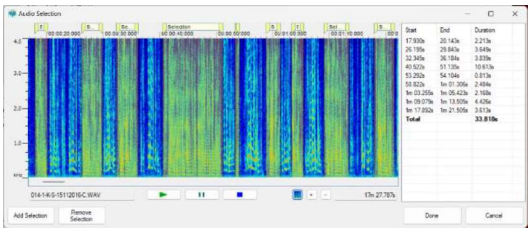

Figure 2. VOCALISE Audio Selections interface with regions selected for comparison.

In addition to selecting subsections of audio directly within VOCALISE, practitioners have the flexibility to import their own timestamps from external analyses, such as Praat TextGrids [6], provided these are appropriately reformatted. This capability enables the focused analysis of specific speech features, such as vowels or other distinctive phonetic or linguistic elements. This capability allows users to investigate the impact of different speech features on speaker recognition analysis.

3.3 Automatic Segmental

Manually creating selections can be very laborious and casework often involves vast quantities of data with limited information. Details such as who and how many people are speaking may not be readily available. Building upon earlier work [7, 8], VOCALISE introduces a novel 'segmental mode' as an automatic alternative for investigating multi-speaker files.

The segmental comparison mode aims to ascertain the presence of a speaker of interest and where their utterances occur within a multi-speaker audio file. This mode segments one set of the audio files submitted for comparison into short, overlapping segments, to be compared against a set of non-segmented, single-speaker audio files containing the speaker of interest.

A score trajectory file generated for each segmented file provides a timeline of the comparison scores for each segment across the file to facilitate further analysis. Higher scores within these segments suggest a greater likelihood of the voice of the speaker of interest being present during those intervals.

VOCALISE allows the forensic practitioner to select between two scoring methods: Max Mode, which outputs the highest score, and Mean Mode, which provides an average score based on all segments above a predefined threshold. Should all regions yield scores below the threshold, the maximum score is returned. Using the spectrogram or waveform, the software displays the highest-scoring region in Max Mode, and all regions above the score threshold in Mean mode. The highlighted regions indicate the location(s) in the recording most likely to contain speech from the speaker of interest. When comparing multiple recordings against each other, the resulting score matrix displays either the highest or average score for each pairwise comparison, depending on the selected scoring method.

The practitioner can activate segmental mode for either analysis or comparison files. They also have the flexibility to define the window size (i.e. the length of the segments) and window slide (i.e. the step size by which a window is moved along the audio file) according to their data specifics. For Mean Mode, a Score Averaging Threshold can be configured to determine the segments contributing to the average score calculation.

Figure 3 shows the VOCALISE window after a segmental comparison, with highlighted regions above a set score threshold in Mean Mode on the spectrogram of a two-minute multi-speaker recording.

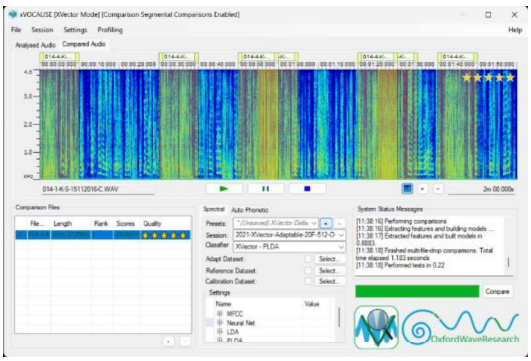

Figure 3. Result view of VOCALISE comparison in Segmental Mode.

Based on the result shown in Figure 3, Figure 4 displays the scores for each segment when compared with the single-speaker audio file, aligned with the waveform view of the segmented audio (these scores are contained within an automatically created trajectory file). The segments highlighted in blue in the waveform exceed the score threshold of -60, indicated by a black horizontal line in the graph below.

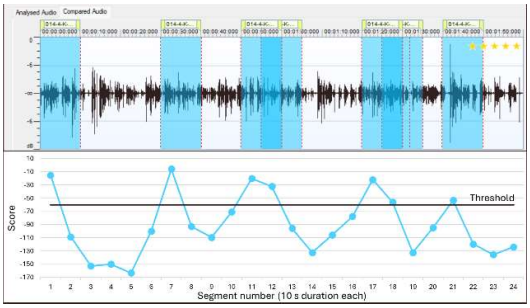

Figure 4. Waveform view of spectrogram from Figure 3 aligned with a graph displaying segmental comparison scores in blue and the specified average score threshold of -60 in black.

The application of Max or Mean Mode serves different casework scenarios. Selecting Max Mode is particularly useful in situations where the primary goal is to find the strongest evidence of a likely match between speakers across the recordings. It highlights just a single segment of speech in a multi-speaker recording: the most likely to contain speech of the speaker of interest. This approach simplifies analysis, which can be advantageous when dealing with a large volume of recordings. Mean Mode offers a more holistic view of speaker similarity across recordings and can aid further investigation by drawing attention to the multiple likely instances of a target speaker, which is particularly useful in long recordings.

3.4 Combining Selections and Segmental Mode

Should the forensic analyst discover that both the recording under investigation (questioned) as well as the reference recording (known) contain speech from multiple speakers, they can combine VOCALISE Selections and Segmental Mode. This means that one side of the comparison will use the predetermined selections for speaker modelling, while the other side will undergo automatic segmental processing.

4 Experiments with a Forensically Relevant Speaker Recognition Database

We consider the two approaches discussed earlier, namely objective quality assessment and multi-speaker handling, using a forensically relevant speaker recognition database called WYRED [9]. We examine the impact of selecting different quality settings and comparison approaches for multi-speaker recordings on speaker discrimination performance, measured using Equal Error Rates (EERs).

4.1 Data Description

This experiment used recordings from the forensically relevant WYRED corpus [9]. This database contains speech from 180 male speakers aged 18-40 years from West Yorkshire, UK. Speakers were recorded over multiple speaking tasks including a staged police interview (Task 1) and an answerphone message to an 'accomplice' (Task 4). The recording duration varies from roughly 13 min to 37 min (Task 1) and about 1 min to 3 min (Task 4). Separate recordings were available for Task 1 interviewer and participant tracks. To create true multi-speaker files, the studio quality tracks were time-aligned using audio fingerprinting in MADCAT [10, 11] and combined to form a single-channel recording. For the single-speaker condition, Task 4 studio quality recordings containing only the participant were chosen. Prior to the automatic speaker comparisons, the multi-speaker Task 1 recordings were trimmed to a duration of 120 s. For the recordings in both tasks, manually determined timing information, which indicates the regions in which the participant is speaking, is available; this enabled the creation of single-speaker Task 1 recordings.

4.2 Quality Variation

To illustrate the impact of audio quality on speaker recognition performance, we present an experiment whereby the quality of samples under comparison is varied systematically. In this example experiment, we consider one dimension of audio quality, namely net-speech duration. For the experiment, we drew samples from the WYRED corpus: specifically, we created a 5-star subset of single-speaker Task 1 (participant only) and Task 4 studio quality recordings, which consisted of 91 speakers, each with a 5-star file in both tasks. Starting with this subset, the net-speech duration of every file was reduced such that a version at each quality rating, 1–5, was created. All quality combinations of Task 1 and Task 4 were then compared using VOCALISE in x-vector mode (with MFCC features and PLDA scoring). The resulting EERs are shown in Table 2.

The results in Table 2 clearly demonstrate the impact of audio quality, specifically net-speech duration, on the resulting (convex hull) EER: The best performance of 1.1% EER is achieved when both files have 5-star net-speech duration (30+ seconds). Performance drops progressively to 9.91% when both files have 1-star net-speech duration (5 seconds). It can also be noted from Table 2 that it is beneficial

even if only one of the files in a comparison is of higher quality.

Star rating (duration)		Task 4				
		5 ★★★★★	4 ★★★★	3 ★★★	2 ★★	1 ★
Task 1	5 ★★★★★	1.10	1.43	2.18	2.5	4.92
	4 ★★★★	1.10	1.56	1.97	2.52	4.63
	3 ★★★	1.78	2.13	2.62	2.80	5.18
	2 ★★	1.98	2.69	2.79	2.90	6.26
	1 ★	4.16	5.20	6.09	5.83	9.91

Table 2. EERs (%) for the comparison of files at varying net-speech durations, from 1 to 5 stars.

4.3 Multi-Speaker Comparisons

Three automatic speaker comparisons were conducted:

1. a naïve comparison of multi-speaker Task 1 recordings of 120 s duration with single-speaker Task 4 recordings,
2. an automatic segmental comparison of multi-speaker Task 1 recordings of 120 s duration with single-speaker Task 4 recordings, and
3. a comparison of single-speaker Task 1 recordings (manually diarised) of 120 s duration with single-speaker Task 4 recordings. Note that the duration of the Task 1 recordings decreases after diarisation.

All comparisons were run using VOCALISE x-vector mode with MFCC features and PLDA scoring. The segmental mode utilised a 10 s window and a 5 s slide and was run in both Max and Mean modes (using the default threshold of -40 in Mean mode).

Table 3 gives an overview of the EERs (convex hull) for each comparison type. It is evident that the highest EER is obtained for the traditional comparison that ignores the multi-speaker recording condition. It was possible to reduce the EER by more than half using segmental mode on the multi-speaker Task 1 recordings. The lowest EER was obtained using the manually diarised Task 1 recordings. We note that the same EER is obtained with Max and Mean mode; this may occur when the discrimination with single-speaker files is very good, as is the case here (the manually diarised EER is 1.44%).

Comparison type	Single-speaker files	Comparison files	EER %
naïve	Task 4 - studio	Task 1 - 120 s - studio	10.01
segmental (max/mean)	Task 4 - studio	Task 1 - 120 s - studio	4.93
manually diarised	Task 4 - studio	Task 1 - 120 s -studio	1.44

Table 3. EERs% for WYRED multi-speaker comparisons.

These results support the use of segmental mode for investigative purposes where swift identification of relevant files is paramount. For the evaluation of evidence, manual diarisation remains the recommended approach.

5 Conclusion

Forensic speaker recognition practitioners must take into account various factors when comparing recordings, including recording quality, duration, and the presence of multiple speakers. One of the problems faced by practitioners is the resource-demanding task of manually editing recordings to isolate single speakers while also ensuring that there is sufficient good quality audio material for comparison. In this paper, we have presented features within the latest version of VOCALISE software that offer pragmatic solutions to these challenges of variable recording quality and multi-speaker recordings.

A star rating based on net-speech duration, SNR, and clipping was introduced as an objective way to measure audio quality. The impact of audio quality variation on speaker recognition performance was illustrated via an experiment with systematically controlled net-speech duration, which showed a decrease in error as star rating increased. In practice, the impact of audio quality variation will be more nuanced, depending on the interplay between quality metrics, e.g. high SNR and short duration versus low SNR and long duration, in addition to the two-sided nature of comparisons. As it is not always feasible to limit automatic speaker recognition comparisons to high quality recordings, star ratings provide an objective measure of audio quality which is especially useful for validation studies.

In response to the challenge of handling recordings containing multiple speakers, a segmental approach

enabling fully automatic comparisons between multi-speaker and single-speaker files was introduced. The capability of the segmental mode in VOCALISE was demonstrated via an experiment that contrasted three approaches to dealing with multi-speaker recordings: naïve comparison, automatic segmentation of the audio, and manual speaker separation. The results indicate that while manual separation achieves the best performance, there is a large relative improvement in performance using the segmental approach compared to a naïve approach. This capability offers great potential for finding speakers of interest within large data collections in time-critical investigations.

The experiments presented here demonstrate pragmatic and effective solutions to common challenges in forensic speaker recognition casework enabled by VOCALISE software.

References

[1] F. Kelly, O. Forth, S. Kent, L. Gerlach, A. Alexander, "Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors," *Proc. AES International Conference 2019*, Paper 27 (2019).

[2] M. I. Mandasari, R. Saeidi, M. McLaren, D. A. van Leeuwen, "Quality measure functions for calibration of speaker recognition systems in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing 21 (11)*, pp. 2425–2438 (2013).

[3] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castan, A. Lawson, "Analysis of Critical Metadata Factors for the Calibration of Speaker Recognition Systems," *Proc. Interspeech 2019*, pp. 4325–4329 (2019).

[4] C. Kim, R. M. Stern, "Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis," *Proc. Interspeech 2008*, pp. 2598-2601 (2008).

[5] A. Atreya, O. Forth, S. Kent, F. Kelly, A. Alexander, "Estimating the Good, the Bad, and the Ugly in Speech Recordings," *Proc. IAFPA Annual Conference 2018*, pp. 64–65 (2018).

[6] P. Boersma, D. Weenink, "Praat: Doing phonetics by computer," [Computer Software], https://www.praat.org/ (1992-2024).

[7] A. Alexander, O. Forth, A. Atreya, S. Kent, F. Kelly, "Not a Lone Voice: Automatically Identifying Speakers in Multi-Speaker Recordings," *Proc. IAFPA Annual Conference 2017*, (2017).

[8] L. Gerlach, F. Kelly, A. Alexander, "One out of many: A sliding window approach to automatic speaker recognition with multi-speaker files," *Proc. IAFPA Annual Conference 2019*, Paper 23 (2019).

[9] E. Gold, S. Ross, K. Earnshaw, "The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework," *Proc. Interspeech 2018*, Paper 0065, pp. 2748–2752 (2018).

[10] Oxford Wave Research, "MADCAT: Multimedia Audio Duplication and Content Analysis Tool," [Computer Software], https://oxfordwaveresearch.com/products/madcat/ (2020).

[11] O. Forth, A. Alexander, "Content Comparison and Analysis (COCOA) of Contemporaneously Recorded Audio Material," *Proc. IAFPA Annual Conference 2014*, pp. 31–32 (2014).