FORENSIC VOICE COMPARISONS IN GERMAN WITH PHONETIC AND AUTOMATIC FEATURES USING VOCALISE SOFTWARE

MICHAEL JESSEN[1], ANIL ALEXANDER[2], AND OSCAR FORTH[2]

[1] *Department of Speaker Identification and Audio Analysis, Bundeskriminalamt, Germany.*
michael.jessen@bka.bund.de
[2]*Oxford Wave Research Ltd, Oxford, United Kingdom*
{anil|oscar}@oxfordwaveresearch.com

In this article, we present a novel forensic speaker recognition system that provides the capability to perform comparisons using both 'traditional' forensic phonetic parameters and 'automatic' spectral features in a semi- or fully automatic way. We evaluate this approach with simulated and real forensic case data in German, which ranges from high quality laboratory audio data to real telephone intercepts. We examine how the forensic expert can use his or her knowledge of the linguistic and phonetic content of the speech and combine it with 'automatic' acoustic analysis of the speech. This approach is shown to provide a level of validation and safeguard against misleading or incorrect identification results. We demonstrate that processing phonetic data will be in many ways complementary and will offer insights into the voice comparison analysis that the classical automatic methods cannot.

INTRODUCTION

In recent years there has been significant academic and commercial research interest in the application of the so called 'automatic' speaker recognition approaches to forensic speaker comparison tasks. These automatic approaches consist of providing audio files from the suspected speaker as well as the questioned recordings or traces to the software system, and the extraction of speaker-specific features and their modeling and comparison is performed by the software. Often this process runs autonomously, and essentially beyond providing the appropriate files, the user has little control or oversight over what happens within the processing.

However, in many countries, the vast majority of forensic speaker comparison casework is performed by forensic phoneticians who have a lot of experience and knowledge in voice comparison and a good understanding of the legal requirements in their area [1]. Many of these experts are currently 'out of the loop' in a fully automatic analysis. They may want to include automatic methods and make their speaker recognition analysis more objective using likelihood ratios and evaluating system performance, but do not have any straightforward means of doing so in a way that meets the necessary requirements for the transparency of such a system.

There is also a requirement for validation and testing. For instance, the automatic systems may depend on the assumptions made by developers, which may not always hold, and the corpora used to develop and test the systems may be different and often not representative of forensic conditions. It is thus useful to both have a convenient way of performing system evaluations of automatic systems that are adapted to the expert's specific casework conditions and to be able to form a second opinion based on phonetics-based information, while keeping constant as much as possible the same modeling techniques and statistical (Bayesian) evaluation framework as in classical automatic speaker recognition.

At the German Bundeskriminalamt (BKA), a system named SPES (*Sprechererkennungssystem* `speaker recognition system') had been developed systematically since 2005 in cooperation with the Technical University in Koblenz (Prof. Broß) in which all the components of the system have been documented and tested [2]. What is not included in SPES is a way of including semi-automatic procedures that are based on acoustic-phonetic information.

We have used a forensic speaker recognition system called 'VOCALISE' (Voice Comparison and Analysis of the Likelihood of Speech Evidence) that provides the capability to perform comparisons using both 'traditional' forensic phonetic parameters and 'automatic' spectral features in a semi- or fully automatic way [3]. 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. Some aspects in the automatic speaker recognition methods of both the BKA-SPES system and of VOCALISE were based upon the development of the system ASPIC (Automatic Speaker Individualisation by Computer) in which the second author has been involved ([4] for details).

1 VOCALISE SPEAKER RECOGNITION SYSTEM

The VOCALISE speaker recognition software 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 was designed to allow the forensic practitioner to statistically model and compare long-term formant information and formant dynamics, along with spectral features like Mel Frequency Cepstral Coefficients (MFCCs) [5, 6]. It provides the capability to perform comparisons using 'automatic' spectral features, 'traditional' forensic phonetic parameters as well as 'user'- provided features.

2 BRIEF DESCRIPTION OF THE VOCALISE USER INTERFACE

In developing VOCALISE, 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 (Fig. 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.

The VOCALISE main page contains a visual display of the waveform of any of the audio and comparison files, which can be selected and played. The playback capability allows to play, pause and thus listen to the files that undergo analysis and there is the capability of zooming and navigating though the signal. The lower left window of the main page shows one or (in this case) several comparison files that can be moved into this window through a simple drag-and-drop action. Also shown in this window is the length of any of the comparison files and the respective likelihood score that is obtained after an analysis file (which can be dropped into the signal display window) has been compared to a comparison file. In order to carry out entire system tests such as n reference recordings compared against m traces, VOCALISE can time-efficiently output result files in CSV (comma separated values) format that contain the results for all the comparisons.

The lower-central section of the VOCALISE main page shows various parameters that can be selected by the user, including the number of Gaussians and the number of features (MFCC coefficients). It also allows the user to select a folder in which the files to be used for training a UBM can be found.

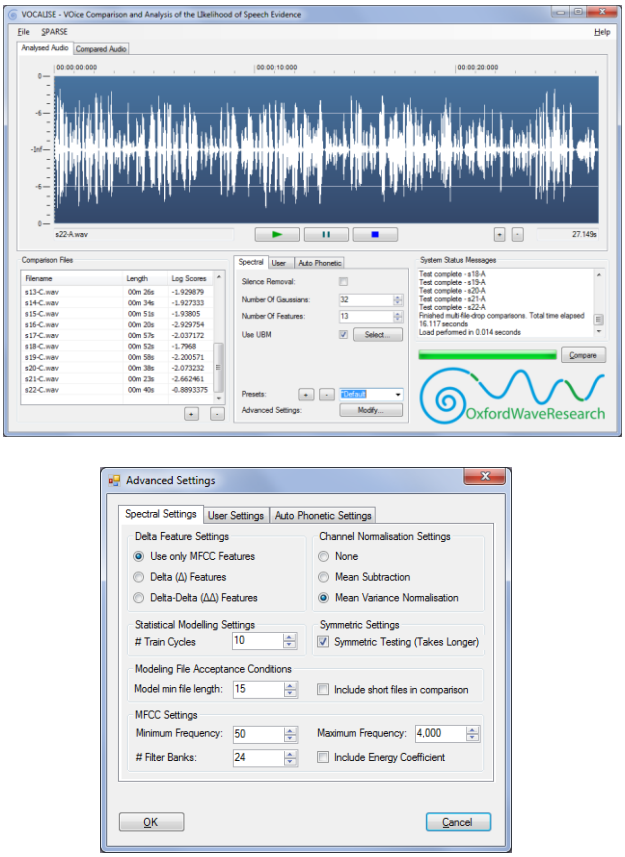

Figure 1: Display of VOCALISE main page (upper) and the Advanced Settings page (lower) during application of the case-data system test (presented in Section 4).

Further analysis parameters can be found on the Advanced Settings page. It includes the option of involving derivatives of the feature vectors like delta and delta-delta coefficients, different channel normalisation methods, different numbers of train cycles for GMM modelling, symmetric testing (inverting the status of training and testing files and taking the average score from both perspectives), and different MFCC settings (frequency range, # of filter banks, inclusion of energy coefficient). There is also the

option of specifying a net duration value below which audio files are not included in the analysis because they are too short (which has to be tested empirically). There are very similar control parameters for each both the autophonetic and user modes. For instance, it is possible to use the derivative 'delta' or 'delta-delta' parameters to phonetic features which will allow the user to model to some extent formant dynamic information. Settings like mean normalisation, applicable to spectral comparisons, are also extended to the two other modes.

3 EXPERIMENTS WITH THE POOL 2010 RESEARCH CORPUS: LONG-TERM FORMANTS

As discussed in the introduction, an important aspect in the design of VOCALISE is its capability of not only being able to carry out established automatic speaker recognition methods but of also being able to process phonetic features such as formant frequencies. This section reports on experiments in which aspects of phonetic data processing with VOCALISE are explored systematically.

3.1 Description of the data

The experiments are based on the research corpus "Pool 2010", which contains laboratory recordings of 100 male adult speakers of German. It was compiled at the Bundeskriminalamt, Germany in order to investigate the intra-speaker and inter-speaker variation characteristics of a number of forensic-phonetic features (see [7, 8] for overview).

Two recordings each from 22 speakers were used, resulting in 22 same-speaker comparisons and 462 different-speaker comparisons. Recordings from 22 other speakers were used to train a UBM. The speakers were adult males and they spoke in a slight regional accent of the West-Central variety of German. The net speech durations (i.e. speech with silences and pauses removed) of the audio files (analysis, comparison, and UBM) ranged from about 20 to 40 seconds. The common factor was the amount of pure vocalic long-term formant material (to be explained below), which was very closely around 10 seconds for each recording. The recordings were in microphone quality but were later transmitted through authentic mobile-phone connections and re-recorded; it is those telephone-transmitted versions that were analysed here. The recordings were not non-contemporary, but they were separated by other recordings and events within a large master recording session. The material from Pool 2010 was intentionally selected in a way that a certain stylistic difference occurred: the speech in the test set was slightly more spontaneous than the one in the training set and in the UBM set. In the test set, subjects spoke freely about their experiences and the observations they made during the recording, whereas in the training set and the UBM set they provided a picture description in which they had to avoid certain key words. When stylistic differences occur in forensic case material, it is often in the same direction, i.e. with the test recording (i.e. the one from the questioned speaker) being more spontaneous than the training recording (i.e. the one from the suspect). However, the amount of stylistic mismatch found in this corpus data is smaller and the overall technical quality much better than what is most commonly found in real forensic cases. Therefore, speaker recognition performance is expected to be much better than in true forensic case material. The evaluations that are carried out based on this material will be referred to as the lab-data system test.

3.2 Spectral Comparisons

First of all, the lab corpus material described here was analysed within the **Spectral** mode of VOCALISE. The features used in the Spectral mode are MFCC (Mel Frequency Cepstral Coefficients) and are modelled using a GMM-UBM approach, which is the case for all the modes in VOCALISE. The analysis settings are those shown in Fig. 1. The EER (Equal Error Rate) obtained from this test was as low as 0.1%. This is a very good result, but it needs to be considered that this result is based on good-quality speech data and might deteriorate in real forensic case data, as will be shown in Section 4. A Tippett-plot of the results is shown in Fig. 2.

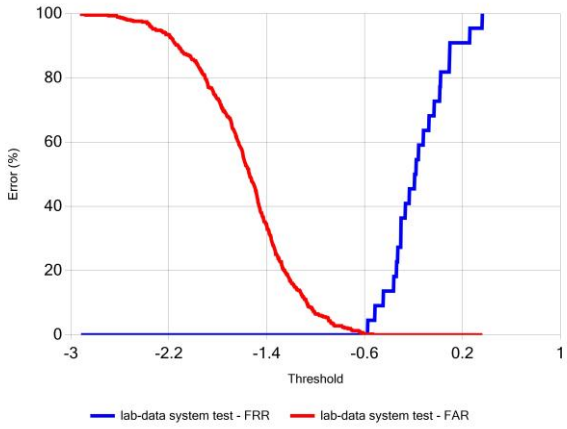

Figure 2: Tippett plot of the lab-data system test using Bio-Metrics software [9]. The x-axis shows the log10-likelihood scores calculated with VOCALISE and the y-axis shows the cumulative proportions of the two distributions. The curve descending from upper left (in red) shows the scores for the different-speaker comparisons, the one ascending towards to the upper right (in blue) shows the scores of the same-speaker comparisons.

3.3 Long-term formant analysis using user-provided features

Secondly, the lab corpus material was analysed with Long-Term Formants (to be abbreviated as LTF). LTF-analysis has been first introduced into forensic phonetics by Nolan & Grigoras [10]. The formant frequencies of different vowel categories have been shown to carry important speaker-specific information [11]. LTF analysis is a particular method of forensic vowel-formant analysis in which formant frequencies are collected during the course of a recording without segmentation into different vowel categories. Since LTF analysis effectively averages across different vowels it is expected to have similar acoustic properties as the central vowel schwa and hence to be an acoustic correlate of individual differences in the length of the vocal tract [12].

In the methodological version of LTF analysis used for the present study, the speech files were labelled, using Praat software [13], in a way that only vowels with visible F1, F2 and F3 (first, second and third formant) are used for the analysis. The total duration of these vowel-only portions of the signal was very closely around 10 seconds for each of the 66 audio files used in the experiments (22 test, 22 training, 22 UBM). These vowel-only portions were concatenated using Praat's Extract function and uploaded into the software Wavesurfer [14]. Within Wavesurfer, formant tracking was applied using Wavesurfer's default settings for formant tracking and any errors by the tracker were corrected manually.[1] The procedure is illustrated in Fig. 3.

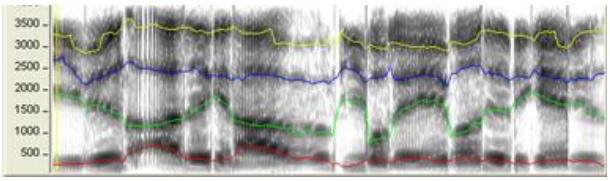

Figure 3: Illustration of a section of an audio recording after concatenated vowel-only portions have been uploaded into Wavesurfer. The display shows a spectrogram (time on x-axis; frequency on y-axis) with overlaid and manually corrected tracks of (from bottom to top) Formant 1 to 4 (F4 is not used in the analysis).

The corrected formant tracks are stored in Wavesurfer as text files with the extension .frm. These text files contain eight columns, the first four containing the formant frequencies F1, F2, F3, F4, and the remaining four containing the formant bandwidths B1 to B4. (F4 and B4 will not be used here due to their close proximity to the upper boundary of the telephone passband.) The mode within VOCALISE that is used for the analysis of the hand-corrected LTF data is called the **User** mode. This is meant to indicate that within this mode of the software, the user has the opportunity to upload any numerical input that is stored in a text file and is arranged in columns. In this analysis, the input originates from formant tracking within Wavesurfer, but in other analyses it could be input from any other software and other phonetic parameters such as fundamental frequency or syllable duration values. In the Advanced Settings (cf. Fig.1), no delta features and no channel normalisation has been applied.

In these tests some of the analysis parameters were varied systematically. One of the parameters that were varied is the number of Gaussians. Whereas there is ample experience with the number of Gaussians necessary in automatic speaker recognition (see [4], including further references], there is so far only limited experience with the GMM-modeling of LTF data [16, 17]. In order to illustrate the situation, Fig. 4 shows an example of the LTF-distribution of the recording of a speaker.

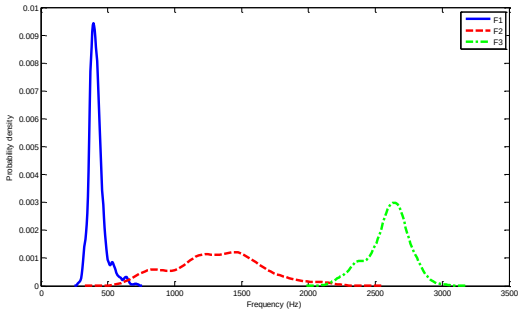

Figure 4: Example of the distributions (probability density functions) of the frequency values of F1 (leftmost, in blue), F2 (center, red) and F3 (rightmost, green) of the recording of a speaker.

In the example shown in Fig. 4, the distributions of formants F1 and F3 are simpler than the one of F2. With F2 it is clear that a single-Gaussian model would not be enough to capture the data, and that a Gaussian Mixture Model (GMM) would be appropriate. For F1and F3, a

[1] Wavesurfer allows for a convenient and time-efficient correction of formant frequency tracking errors. Praat does not have an option for correcting formant tracks, otherwise the entire analysis would have stayed within Praat and data transfer between software packages would have been unnecessary. Some difficulties with the Wavesurfer formant tracker have been reported when formants are close to each other [15], but manual correction should be able to accommodate such problems. Wavesurfer does not provide a graphical display of the formant bandwidths (to be explained later in this text) and therefore no manual correction of potential errors in bandwidth determination was carried out.

single-Gaussian model might be more adequate than for F2, but the distributions of F1 and F3 shown here are not entirely regular either and could also benefit from modeling with GMM. Furthermore, there are examples of speakers (not shown) where F1 and F3 have a more complex distribution than the ones shown here. Based on examples like this, the hypothesis emerges that Gaussian Mixture Modeling, as opposed to modeling with single Gaussians, is of benefit for the voice comparison process. If, according to this hypothesis, more than one Gaussian is useful, it has to be studied how many Gaussians are necessary. Perhaps in order to capture distributions such the one of F2 shown in Fig. 4 not very many Gaussians are required. Becker et al. [17], based on a section of Pool 2010 different from the one here, reports that 8 Gaussians are necessary to obtain the best results, but perhaps the full performance is reached at even lower values.[2]

Another parameter that was varied is the inclusion or non-inclusion of the formant bandwidths. Formant bandwidths can be measured manually "by noting the frequencies that are 3 dB below the frequency with the maximum amplitude [12, p. 88]", but formant bandwidths can also be determined automatically within LPC analysis (as it is performed here using Wavesurfer), which might not yield exactly the same results. In some of the early studies on speaker identification, formant bandwidths have been shown to carry some speaker-discriminative information [16]. Becker et al. [16] showed some improvement of speaker recognition performance if the bandwidths were included (based on the same data as in Becker et al. [17]).

In the experiments reported here, the number of Gaussians was varied from 1 to 8. This series of eight different numbers of Gaussians was tested in two sets, one based on F1, F2, F3 and one based on F1, F2, F3 plus B1, B2, B3, i.e. the first set contained the frequencies of the first three formants, whereas the second set also contained their bandwidths. The results expressed in terms of EER are shown in Fig. 5.

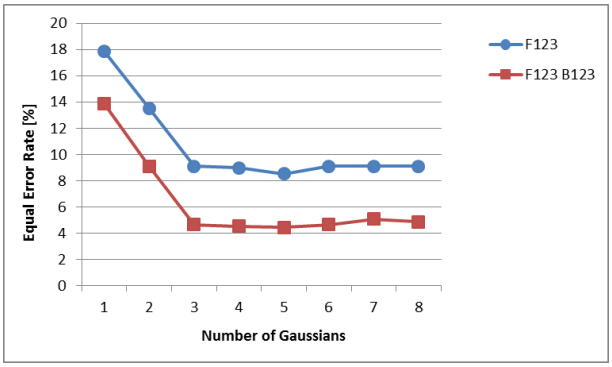

Figure 5: Results of the lab-data system test with the User mode (hand-corrected Long-Term Formants). Results in terms of Equal Error Rate (EER, y-axis), when the number of Gaussians is varied from 1 to 8 (x-axis) in two series, one with F1, F2, F3 (circles, blue) and one also including the bandwidths (squares, red).

The results shown in Fig. 5 are straightforward. Firstly, every step-by-step increase from one to three Gaussians leads to an improvement of speaker recognition performance, but from three Gaussians on, there is practically no improvement. Hence, three Gaussians are sufficient to model the LTF data for voice comparison purposes. Secondly, including the bandwidths increases performance. There is fairly little interaction with the number-of-Gaussians parameter, i.e. a similar improvement occurs for all number-of-Gaussians.

3.4 Long-term formant analysis using VOCALISE's automatic phonetic feature extraction

So far, the lab-data system test (i.e. data based on the Pool 2010 corpus) was carried out with the Spectral mode (automatic speaker recognition) and the User mode (LTF analysis based on manually corrected formants). It was also carried out with a third mode of VOCALISE, which is called the **Autophonetic** mode. The intention behind the Autophonetic mode is to offer automatic feature extraction methods for features that belong to the domain of acoustic phonetics and that are used routinely by forensic phoneticians. Currently, the Autophonetic mode offers the capability for LTF analysis, but further developments are possible. Fig. 6 shows the main page of the user interface for the Autophonetic mode. The difference to Fig. 1 lies in the part where boxes can be checked for the formants that are intended to be included in the analysis. The technology behind the automatic formant extraction is based on the Praat software. The formant bandwidths are currently not included.

[2] It should be mentioned that Fig. 4 simplifies the situation insofar as the actual modeling of the formant data in VOCALISE is *multivariate*, i.e. each feature vector is situated in three-dimensional space when F1, F2, F3 are used or in six-dimensional space when also the bandwidths are included. What is shown in Fig. 4, instead, are three univariate distributions.

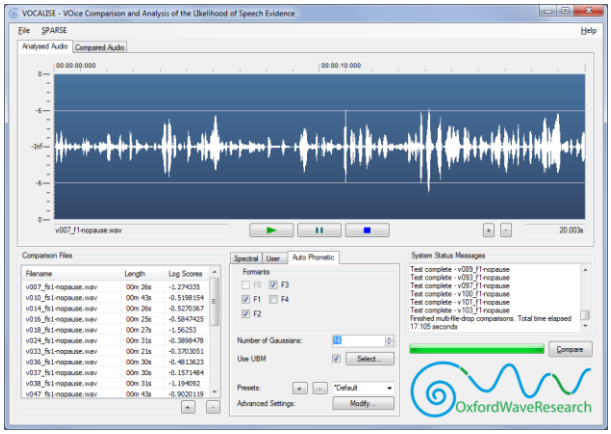

Figure 6: Display of VOCALISE main page during application of the lab-data system test with the Autophonetic mode (automatic LTF-analysis).

Using the Autophonetic mode, a variety of different system tests were carried out. As shown in Fig. 6, the Formant frequencies F1, F2, F3 were used as features. The number of Gaussians was varied systematically as in the LTF tests before, but this time the number was increased up to 16. The remaining settings are the same as in the User mode. The results are shown in Fig. 7.

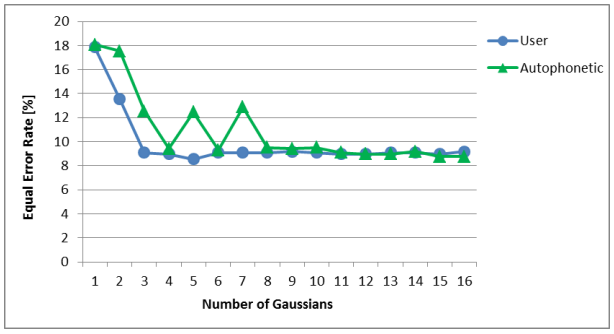

Figure 7: Results of the lab-data system test based on the formant frequencies F1, F2, F3. Results from the Autophonetic mode (automatically extracted Long-Term Formants; shown with green triangles). This is compared to the results from the User mode (hand-corrected Long-Term Formants) repeated from Fig. 5 and supplemented with tests from 9 to 16 Gaussians.

The results in Fig. 7 show that the Autophonetic mode needs more number of Gaussians until the performance stabilizes. With the User mode stabilization occurred at three Gaussians, here it is at eight. However, beyond that point, the performance of the Autophonetic mode is practically the same as the one of the User mode. What is interesting about that result is, first, that two fairly different ways of arriving at Long-Term Formants yield such similar outcome in terms of EER. Secondly, the results show that a fully automated system for measuring the formant frequencies gives the same speaker recognition result as a system that is based on manually selecting vocalic speech portions and correcting formant tracks. Whether this result carries over to the handling of real case data remains to be studied in the future. The advantage of the User mode that stability is reached with lower numbers of Gaussians compared to the Autophonetic mode is only marginally relevant.

4 EXPERIMENTS WITH REAL CASE DATA

The VOCALISE system was used in analysing anonymised case data collected from telephone interception recordings in Germany. This will be referred to as the case-data system test. Two non-contemporary recordings each from 22 speakers were used, which provided 22 same-speaker comparisons and 462 different-speaker comparisons. The speakers were adult males and speaking German, some of whom had regional or ethnic accent. The net speech durations of the audio files analysed ranged from about 20 to 60 seconds. The speech style in the recordings was completely natural and the recording conditions were largely similar with no significantly discernible difference in channel distortion. It is reasonable to consider samples to come from broadly matching conditions. Nevertheless, each recording contained speech of varying degrees of vocal loudness, emotionality, background noise, distortions, etc., making the voice comparisons a challenging and a fully 'forensically realistic' task. In addition to the 22 training recordings (referred to as comparison files in the VOCALISE software) and the 22 test recordings (referred to as analysis files), natural telephone recordings from 25 other male German speakers of about one minute net duration each were used to train a UBM (Universal Background Model).

These data were analysed with the Spectral mode of VOCALISE. Exactly the same analysis settings were used here as in the lab-data system test, i.e. those shown in Fig. 1. Results are shown in Fig. 8.

As indicated in the DET plot and detectible from the Tippett plot (intersection between same-speaker and different-speaker curves), the EER of the case-data system test was at 11.3%. This is a reasonable value compared to other system tests on telephone-interception-type forensic material using systems based upon the use of MFCC features within a GMM-UBM approach [19].[3]

[3] As can be seen in the Tippett plot here or in Fig. 2, the scores are not yet calibrated. Calibration is possible, for example within in the Bio-Metrics software, if further development data are provided or if a cross-validation procedure is used ([20] for illustration of the methodology).

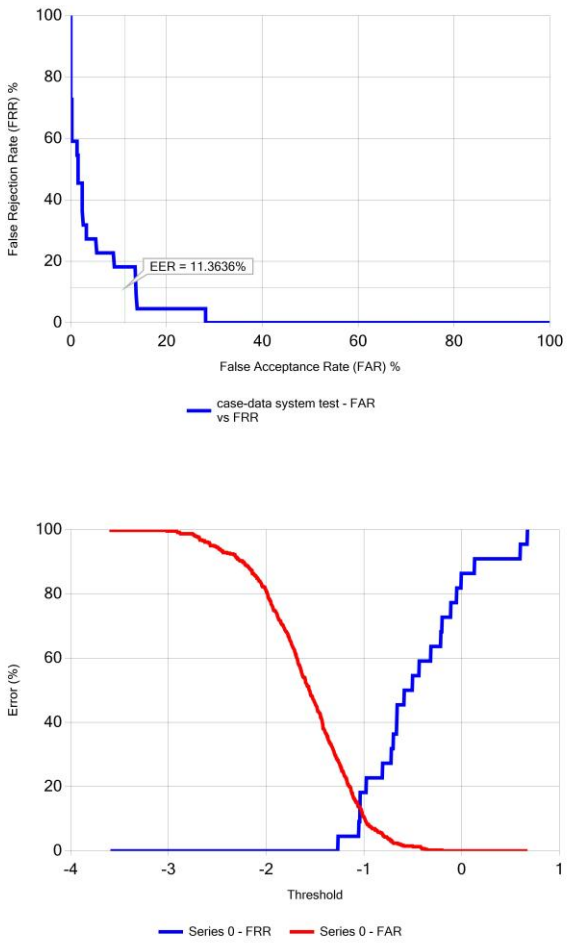

Figure 8: DET plot (upper) and Tippett plot (lower) of the case-data system test using Bio-Metrics software. In the DET-plot, false identification rate is plotted on the x-axis and false rejection rate on the y-axis. Tippett plots were explained in Fig. 2.

5 SELECTIVE PROCESSING (SPARSE)

In its recent development, VOCALISE has been extended by a component called SPARSE (Selective Processing of Annotated Regions of Speech Efficiently). When SPARSE is enabled, the spectral, user, and auto phonetic methods described above are region-conditioned to speech sounds, speech styles or other subsections of recordings that are labeled by the user. These labels in the form of Praat TextGrids or other formats are recognised by VOCALISE and all training and testing is limited to the regions that are of interest.

As a first test, SPARSE has been applied to the speech data and user-provided formant measurements described in Section 3.3. In contrast to the methods in Section 3.3, the formants were not extracted across vowels but were limited to hand-labeled vowels. In the speech data, the three vowels (in SAMPA notation) /a/ (short/lax a-vowel), /I/ (short/lax i-vowel) and /@/ (schwa) occurred most frequently, therefore SPARSE analysis concentrated on these vowel categories. Using F1, F2, and F3 and varying the number of Gaussians, the results showed that EER for the /I/-regions was on average slightly above 20%, for the /a/-regions it was slightly below 20%, and for the /@/-regions was slightly above 40%. The poor results for schwa are likely to be the result of the high degree of coarticulation – and hence increased intra-individual variation – experienced by this sound. The results also indicate that the marked decline of EER values from one to three Gaussians shown in Figs. 5 and 7 does not apply to single vowels, which suggests that single vowel categories can quite adequately be modelled with single Gaussians. Fig. 9 shows the user interface of VOCALISE in these experiments, which will be extended and more fully documented in future research.

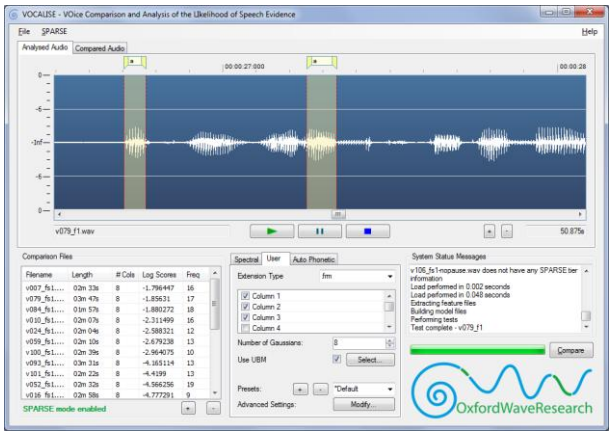

Figure 9: Display of VOCALISE main page during application of the SPARSE component. The Waveform display shows two tokens of /a/ which had been labeled in Praat and are automatically imported into VOCALISE by SPARSE.

6 CONCLUSIONS

VOCALISE makes it possible to apply classical automatic speaker recognition transparently and 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, mostly data-driven and less directly connected to the theory of speech production [11]. 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

[1] E. Gold & P. French "International practices in forensic speaker comparison" *The International Journal of Speech, Language and the Law* vol. 18, pp. 293—307 (2005).

[2] T. Becker, Michael Jessen, S. Alsbach, F. Broß and T. Meier, "The BKA Forensic Automatic Voice Comparison System", Proceedings of *Odyssey: The Speaker and Language Recognition Workshop*, pp. 58—62 (2010).

[3] http://www.oxfordwaveresearch.com/j2/products/vocalise

[4] A. Alexander, *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*, Ph.D. Dissertation, EPFL Lausanne (2005).

[5] Marianne Jessen, O. Forth & 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).

[6] A. Alexander, O. Forth, Marianne Jessen, Michael Jessen, „Speaker recognition with Phonetic and Automatic Features using VOCALISE software", Paper presented at the 22th *Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Tampa (2013).

[7] Michael Jessen, O. Köster, S. Gfroerer, "Influence of vocal effort on average and variability of fundamental frequency" *The International Journal of Speech, Language and the Law* vol. 12, pp. 174—213 (2005).

[8] Michael Jessen, *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*, LINCOM EUROPA (2012).

[9] http://www.oxfordwaveresearch.com/j2/products/bio-metrics-1-1-performance-metrics-software-2

[10] F. Nolan & C. Grigoras, "A case for formant analysis in forensic speaker identification" *The International Journal of Speech, Language and the Law* vol. 12, pp. 143—173 (2005).

[11] Rose, P., *Forensic Speaker Identification*, London: Taylor & Francis (2002).

[12] Ladefoged, P., *Elements of Acoustic Phonetics, 2. Ed.*, Chicago: The University of Chicago Press (1996).

[13] http://www.fon.hum.uva.nl/praat

[14] https://www.speech.kth.se/wavesurfer

[15] N.F. Chen, W. Shen, J. Campbell & R. Schwartz, "Large-scale analysis of formant frequency estimation variability in conversational telephone speech *Proceedings of INTERSPEECH* (Brighton), pp. 2203—2206 (2009).

[16] T. Becker, Michael Jessen & C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models *Proceedings of INTERSPEECH* (Brisbane), pp. 1505—1508 (2008).

[17] T. Becker, Michael Jessen & C. Grigoras, "Speaker verification based on formants using Gaussian mixture models" *Proceedings of NAG/DAGA*, (Rotterdam), pp. 1640—1643 (2009).

[18] K.N. Stevens, "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds*" Proceedings of the International Congress of Phonetic Sciences* (Montreal), pp. 206—232 (1971).

[19] D.A. van Leeuwen et al., "NIST and NFI-TNO evaluations of automatic speaker recognition" *Computer Speech and Language* vol. 20, pp. 128—158 (2006).

[20] G.S. Morrison, "A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model - universal background model (GMM-UBM)" *Speech Communication* vol. 53, pp. 242—256 (2011).