Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

SPES: The BKA Forensic Automatic Voice Comparison System

Timo Becker[1], *Michael Jessen*[1], *Sebastian Alsbach*[2], *Franz Broß*[2], *Torsten Meier*[2]

[1]Federal Criminal Police Office, Germany
[2]University of Applied Sciences Koblenz, Germany

timo.becker@bka.bund.de, michael.jessen@bka.bund.de
salsbach@arcor.de, bross-franz@t-online.de, tmeier@fh-koblenz.de

Abstract

The BKA voice comparison system SPES is designed for forensic examination of speech recordings. The classical GMM-UBM framework based on MAP adaptation as described by Reynolds et al. [1] is extended by the generation of recording adapted background models (RABMs). We present results from experiments using real case data. These results show how the most critical properties of real case recordings such as duration, channel, and samples per speaker influence system performance.

Index Terms: voice comparison, Gaussian Mixture Models, adaptation, forensic, speaker recognition

1. Introduction

The automatic voice comparison system's name SPES originates from the German term ”SPrecher-Erkennungs-System” which means ”speaker recognition system”. The system is developed in cooperation with the Federal Criminal Police Office of Germany (Bundeskriminalamt – BKA), the University of Applied Sciences Koblenz, and the Department of Phonetics, University of Trier, Germany. This is achieved using anonymised real case data which mainly originate from the Federal Criminal Police Office. Both development and testing is performed accounting for realistic conditions concerning real casework. The current system's design and the parameter settings result from years of development under this cooperation. Hence, the SPES system design might differ from non-forensic state-of-the-art speaker recognition systems. Here we provide a description of the system, focusing on these differences.

To estimate the SPES system performance, evaluations using realistic data have to be run. Speaker recognition evaluations like the NIST speaker recognition evaluations [2] might not give us appropriate performance measures because here, usually, a totally blind processing of speech samples is required. This is not the case for forensic applications, where blind processing might be a disadvantage, for example when introducing errors from automatic speech detection algorithms or automatic speech recognition. Here, results can be improved by involving experts in the process. Additionally, some tasks like speech detection still can be done more accurately by manual labeling. Therefore, manual processing is preferred whenever it is expected to achieve more accurate and reliable results.

Another difference between automatic speaker recognition evaluations and speech sample comparisons in the BKA's special forensic setting lies in the recordings' properties. As we will explain later, we usually have to deal with data from different channels, with different speaking styles, with different emotional states, with different acoustic environment properties, and different durations (see Ramos et al. [3] and Fauve et al. [4] for a discussion of those factors and the resulting problems for forensic automatic speaker recognition). Additionally, the number of recordings per speaker differs. All those influencing factors require adaption of system parameters to this special setting. As a consequence, the BKA forensic automatic voice comparison system cannot easily be used for standard automatic speaker recognition evaluations. Because of this specialisation, evaluation results based on non-forensic recordings cannot give us insight into our system's performance for our special forensic setting. In this paper, we will emphasize the most relevant settings and the resulting system specifications.

2. System Description

Each case consists of a number of suspect, offender and impostor recordings. The assignment to these categories is usually provided by the police. An expert listens to the recordings and decides if processing by SPES can be done. When passed on to SPES, every recording has to meet the following requirements:

- minimum 8 KHz sampling rate
- 16 Bit word length
- linear PCM Quantisation
- no voice disguise

The recordings are manually labeled. The labeling includes specifying the following attributes:

- speaker sex
- language spoken
- speaking style (spontaneous, reading, lombard, etc.)
- transmission channel (wiretapping, studio or telephone recording)
- clipping
- distortions
- manipulation (e. g. pitch manipulation)

The expert manually detects speech pauses and strong interferences and deletes them. After that, the speech signal is automatically transformed to a sampling rate of 8 kHz and it is bandpass filtered (300-3400 Hz). Then *relative spectral transform - perceptual linear prediction cepstral coefficients (RASTA-PLPCC)* [5, 6, 7] are computed, Δ features attached. A feature vector mapping[8] is not applied, because it did not increase the system's performance on the real case BKA data significantly. This might be due to the manifold and often unclear channel properties of our real case data.

SPES is basically a UBM-GMM system as described by Reynolds et al. [1]. However, it turned out that system performance depends more on the data used than system design or parameters. Therefore, the single UBM is extended by *recording adapted background models (RABMs) [9]* which create different background models for each suspect.

Based on a population P_{UBM} which consists of a collection of 1383 real case recordings, a standard UBM with 2048 mixtures is computed using *expectation maximisation (EM)* [10]. For a distinct population of real case data $P_0 \cap P_{UBM} = \emptyset$ with N_0 speakers, models are generated using MAP adaptation. P_0 has to contain only recordings of speakers which have the same recording conditions as the suspect recording.

The following steps are done for each suspect recording $s \in P_{Ref}$, where $P_{Ref} \cap P_0 = \emptyset$ and $P_{Ref} \cap P_{UBM} = \emptyset$:

1. Compute likelihood ratios of all N_0 models of P_0.
2. Based on these likelihood ratios, select the N_{max} most similar recordings from P_0. Use a threshold θ.[1]
3. Pool these recordings together and generate an RABM Λ_s with 600 mixtures by using EM.
4. Generate a GMM λ_s via MAP adaptation from Λ_s.
5. Compute likelihood ratios based on λ_s and Λ_s.

Note that the RABM creation significantly increases computation time. SPES supports multi-core processing which reduces the increased computational cost to an acceptable minimum. At the moment, processing of a case using about 5 recordings takes about 4 hours (including human interaction).

For likelihood ratio calculation, SPES uses both the direct and the scoring method as described by Alexander and Drygajlo [11]. In forensic casework, we observed that the scoring method always produces lower error probabilities and adequate likelihood ratios. However, in many cases, the condition that enough recordings of the suspect speaker exist to estimate intra speaker variability is not met. Here, we only use the likelihood ratio scores as described in step 5, keeping in mind that usually an application of the scoring method leads to a better result.

The likelihood ratio scores are finally made symmetrical. This is achieved by computing λ_s for both suspect and offender recordings. Then the offender feature vectors are tested against the suspect model and vice versa. By doing this, two likelihood ratio scores for speech sample comparison are computed. We found out that selecting the maximum of the two scores increases system performance and gives more reliable results. An explanation for this might be, that often disturbances or signal artifacts cause outliers in the score distribution.

The system finally returns a calibrated [12] likelihood ratio score. Since we here focus on the discrimination task which is not affected by calibration, we omitted the calibration step in our experiments.

3. Experiments

The development of SPES is based on a corpus of currently

- 564 male and 25 female speakers out of about 2100 recordings from real cases and
- 303 male and 18 female speakers out of 321 recordings from emergency calls.

We use these recordings because previous experiments revealed a strong difference of performance between artificially created corpora and real case data. Additionally, when producing results for real cases, for the sake of reliability, estimations of error probabilities have to originate on statistics based on comparable data. The following experiments are based on a selection of 747 recordings from 182 speakers.

The assignment of recordings to speakers is based on exterior knowledge about the case as well as manual phonetic-acoustic analysis. We are aware that this might include errors. In case of assignment errors, we expect the automatic system to rather do the correct assignment than repeat the wrong assignment. This would result in an increase of the error rate. So when we evaluate our system and provide the results to the court (maybe including unknown assignment errors), we have a careful and conservative estimation of system performance.

3.1. Channel

In previous experiments, we have seen that the speaking style as well as the transmission channel have great influence on system performance. Hence, we chose to specify these attributes (see section 2) for each recording. Most of the recordings include spontaneous speech which is transmitted over the telephone. Therefore, we will provide results for *non-restricted (nr)* recordings (all 747 recordings) and recordings from the restricted channel condition *telephone spontaneous (ts)*. About 81% of the recordings belong to the ts condition (99% of the offender recordings, 56% of the suspect recordings). Note that we are not able to use the $\neg ts$ condition only because of too few data for calculating reliable error probabilities. The following results for duration and samples per speaker will be shown for both nr and ts conditions.

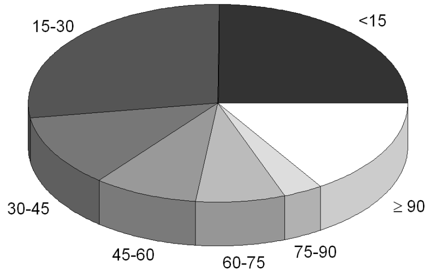

Figure 1: Net durations in seconds for 747 recordings

[1] In previous experiments we found that speaker recordings which have strong noises present in the signal show a high similarity to recordings of other speakers with similar noises. The usage of a threshold compensates for this effect.

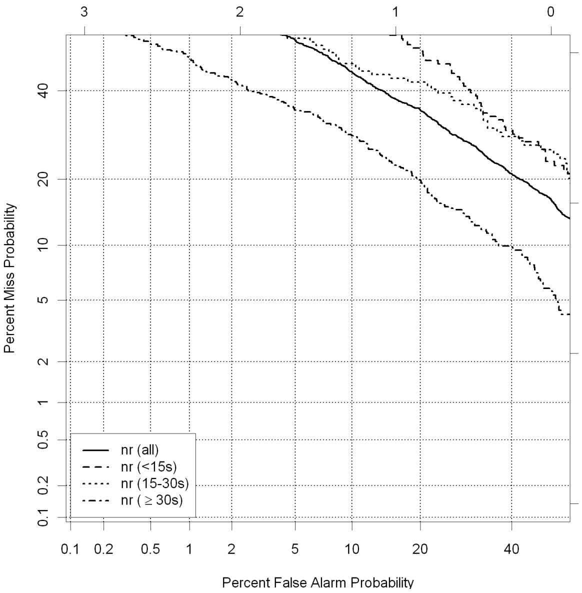

Figure 2: DET-Plot [14] for different durations (nr condition)

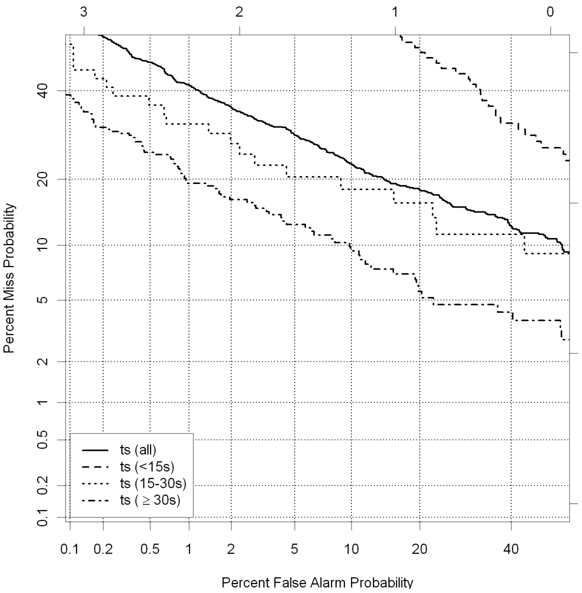

Figure 3: DET-Plot for different durations (ts condition)

3.2. Duration

See figure 1 for a distribution of the net durations of 747 recordings from real cases of the BKA. Note that about a quarter of the recordings is shorter than 15 seconds and the majority of recordings is shorter than 30 seconds. However, suspect recordings tend to be longer (ca. 38% shorter than 30 seconds) than offender recordings (ca. 63% shorter than 30 seconds). It is known that short recording signal durations cause significant performance losses in automatic speaker recognition [4, 13]. This motivated us to adapt to this sparse data as best as possible. The RABM generation and symmetrising the likelihood ratio scores are the most important consequences of this.

See figures 2 and 3 for system performance results for using different durations. Figure 2 shows results for the nr condition, figure 3 shows the results for the ts conditions. It can be seen that generally the ts condition gives better results than the nr condition. Also, system performance is worst when using recordings which are shorter than 15 seconds while system performance is best when using recordings which are longer than 30 seconds. This applies to both the nr and ts condition. When using recordings of all durations, system performance lies between those two extremes. Concerning the recordings which have durations between 15 and 30 seconds, results are more complicated. Although system performance for using recordings of this time interval is better for the ts condition than for the nr condition, those results have to be treated carefully because performance estimation for the ts condition is based on few recordings and might thus not be adequate.

3.3. Samples per Speaker

In many cases, there exists more than one recording for one speaker[2]. See figure 4 for the distribution of samples per speaker. About 55% of the speakers have only one recording, the remaining 45% at least two recordings.

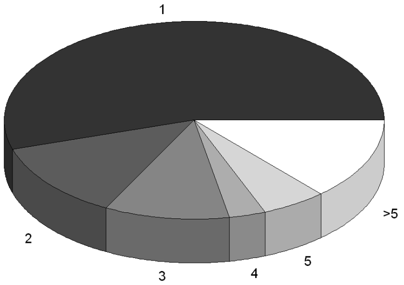

Figure 4: Recordings per speaker

We found that using different methods of score selection improves system performance. See figures 5 and 6 for speaker recognition tests using four types of score selection (figure 5 shows the nr conditions while figure 6 shows the ts condition):

1. *single*: all recordings are compared pairwise
2. *max(suspect)*: select the maximum score of all suspect recordings
3. *max(offender)*: select the maximum score of all offender recordings
4. *max(suspect, offender)*: select the maximum score of all recordings

When looking at figures 5 and 6, it can be seen again, that generally the ts condition gives better results than the nr condition. For both conditions, selecting single scores performs worst. Selecting the maximum of the suspect scores gives better results.

[2] Note, that we rely on exterior information about assignments of recordings to suspects and offenders.

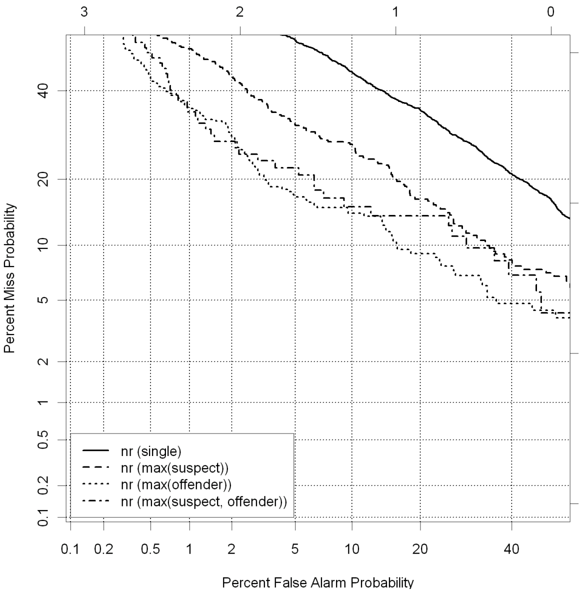

Figure 5: DET-Plot for different selection of maximum likelihood ratio scores (nr condition)

The best performance results from selecting the maximum of either the offender or both scores. We neglect the differences of the lower right corner in figure 6 because the prediction in this area is unreliable due to too few data.

4. Discussion

The results show that using the ts condition gives better results than using all available data. Also, using extremely short recordings correlates with a higher error probability. The degradation of automatic speaker recognition performance in channel mismatch conditions as well as duration is well known [15, 2]. However, we also show that using the maximum score of one speaker improves system performance (given that more than one recording is available). The best performance results from selecting the maximum scores of the offender or both suspect and offender scores. An explanation for the similar results might be that offender recordings tend to be more heterogeneous than suspect recordings and hence benefit from maximum likelihood ratio score selection.

From these results we draw the following conclusions for our casework:

- To minimise the duration effect, we manually delete pauses and distortions in order not to discard any useful data. We also do not use recordings with a net duration shorter than 15 seconds.
- We use all available recording data from one speaker and select the maximum score.
- In any case, we evaluate our system using only comparable data (especially concerning transmission channel and speaking style) in order to make a reliable prediction of error rates.[3]

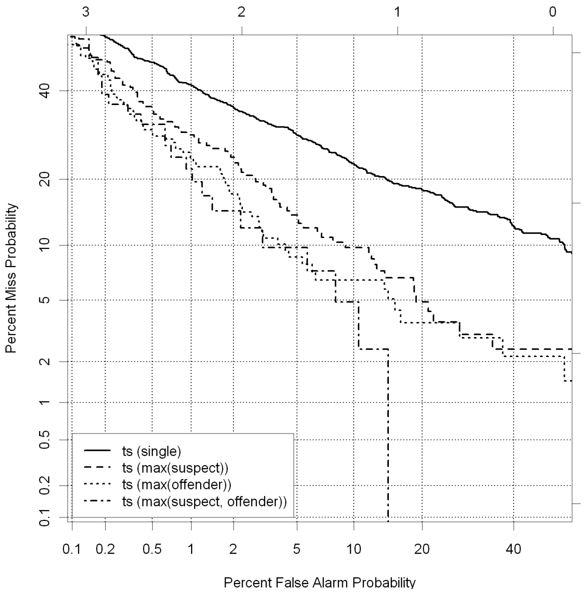

Figure 6: DET-Plot for different selection of maximum likelihood ratio scores (ts condition)

Having seen the different performances of one system under different real case conditions, we emphasize that there is no single measure for an automatic voice comparison system which generally can express system performance. For each case, system performance has to be specified regarding to the case's recording and speakers' conditions. These conditions can help to improve reliability of the automatic voice comparison system when compiling the training corpus which is used to estimate the error probability.

After examination of a case's recordings, a corpus with the same attributes has to be selected from past real case recordings. Then a full evaluation has to be conducted, results have to be provided to account for error rates. We propose to present the case's likelihood ratio score together with the corresponding DET-Plots and APE-Plots [12]. This enables triers to estimate both the similarity of the voices as well as the error probability. Our results show that system performance evaluation is heavily affected by channel mismatch, duration mismatch and mismatch concerning samples per speaker. Standardised evaluations can only be used to estimate system performance for the data used in the evaluation and data with comparable attributes. We emphasize that performance results obtained from evaluations under conditions which differ from real case recordings must not be used to estimate error probabilities for results of real case analyses.

5. Outlook

We have presented results for not using the scoring method. When looking at more than 2000 recordings from 226 real cases of the BKA, only 1% of those cases exactly comply with the necesarry conditions concerning the scoring method. About 22% of the cases nearly comply with the conditions. It might

[3] This is done together with a system calibration as described by Brümmer and du Preez [12].

help to improve system performance and reliability if we modify the scoring method in a way that it can be used on more of our real case data. Also, the application of calibration has to be included in such a modification.

Investigation of other modeling methods such as Support Vector Machines (SVM) and other session mismatch normalisation techniques applied to our real case data might also be interesting. We did not focus on those yet because we do not know how big the influence of our real case data's conditions on their system performance is. This is supported by the fact, that session mismatch normalisation techniques such as Factor Analysis (FA) and Nuisance Attribute Projection (NAP) are sensitive to speech duration [13].

In the future we will continue to expand our real case data collection in order to account for more influencing factors. Only large corpora of recordings under different conditions enable us to give meaningful results and reliable error probability estimations.

6. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora – 2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.

[3] D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, and J. J. Lucena-Molina, "Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish," in *Proceedings of Interspeech 2008 incorporating SST'08*, 2008, pp. 1493–1496.

[4] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification," in *Proc. of Odyssey*, Stellenbosch, South Africa, 2008.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[6] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP Speech Analysis," Tech. Rep. 92-069, 1991.

[7] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[8] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Hong Kong, 2003, pp. II–53–56.

[9] T. Becker, M. Jessen, S. Alsbach, F. Broß, and T. Meier, "Automatic Forensic Voice Comparison Using Recording Adapted Background Models," in *Proceedings of the AES 39th International Conference on Audio Forensics*, Hillerød, Denmark, 2010, to be published.

[10] T. K. Moon, "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[11] A. Alexander and A. Drygajlo, "Scoring and Direct Methods for the Interpretation of Evidence in Forensic Speaker Recognition," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP)*, 2004.

[12] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Lanugage*, vol. 20, pp. 230–275, 2006.

[13] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic Speaker Recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, March 2009.

[14] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of EUROSPEECH'97 - 5th European Conference on Speech Communication and Technologys*, Rhodes, 1997, pp. 1895–1898.

[15] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification," Ph.D. dissertation, Georgia Institute of Technology, 1992.