Pass Phrase Based Speaker Recognition for Authentication

Heinz Hertlein[†], Dr. Robert Frischholz[†], Dr. Elmar Nöth*

[†] HumanScan GmbH
Wetterkreuz 19a
91058 Erlangen/Tennenlohe, Germany

* Chair for Pattern Recognition, Computer Science 5
FAU University of Erlangen-Nuremberg
Martensstr. 3
91058 Erlangen, Germany

{h.hertlein,r.frischholz}@humanscan.de
noeth@informatik.uni-erlangen.de

Abstract: Speaker recognition in applications of our daily lives is not yet in widespread use. In order for biometric technology to make sense for real-world authentication applications and be accepted by end users, convenience of use, robustness and accuracy of such a system are equally important. This paper defines these requirements for pass phrase based voice authentication embedded within a multi modal biometric system and describes methods and algorithms developed and optimized for the demands of such an application. Classification is based on dynamic time warping which can cope with limited training data. MFCC features which have been optimized for speaker specific properties are used. Robustness of the system is increased with speech enhancement and cepstral mean subtraction. Furthermore, vector quantization with speaker specific codebooks is applied in order to decrease storage requirements for the biometric template. On an appropriate data base, a verification EER of 2.7% is achieved with limited training and test material.

1 Introduction: Speaker recognition in a multi modal system for authentication

The BioID SDK (Software Development Kit) [Bio00] is a multi modal, biometric system for authentication, which enables integrators to use advanced and highly specialized biometric technology in end user products without expert knowledge about sophisticated pattern recognition algorithms. BioID is based on a combination of face, lip movement and voice. The simultaneous use of multiple biometric traits achieves good recognition accuracy, although the effort for the user in training as well as to be authenticated is small.

This article is dedicated to techniques and algorithms of the voice trait in particular. In chapter 3, for example, state of the art classification techniques of automatic speaker

recognition are described and compared, especially in regard to the actual real-world requirements of the BioID system. Although it should be noted that voice authentication is embedded within a multi modal system, the relevant techniques are described largely independent of the other biometric traits. Crucial for selection and optimization of the algorithms for speaker recognition in BioID are the particular requirements, which will be elaborated on in the next chapter.

2 Pass phrase based voice authentication and its requirements

The voice trait of authentication with BioID is based on a pass phrase that is chosen by the user and that has to be uttered in order to be granted access. Typically, the name of the person is used as pass phrase. In order to increase security, however, an arbitrary phrase that is to be kept secret can be used. As with any pattern recognition problem, a distinction can be drawn between the training phase, in which the system "learns" relevant classes using sample patterns, and the recognition process, which purpose is to classify an unknown test pattern.

2.1 Limited utterance lengths

Accepted are recordings of utterances which contain at least 400 ms of speech, detected automatically by the system. The duration of all recordings is exactly one second. In training, at least five recordings are required by the system in order to be enrolled. For a classification decision, a single utterance of the pass phrase is sufficient. This applies for identification, where the most similar person known from training is determined, as well as for verification, where an identity claim is to be affirmed or rejected by the system. So, the requirement to be able to work with as little as five seconds of training data for model building and one second test speech for recognition has to be accounted for in regard to the algorithmic realization.

2.2 Model size

In addition to the need to be able to cope with limited amount of sensor data, the storage requirement for the user model (also called *biometric template*) is an important issue as well. For example, there are applications which store the user template on smart cards or iButtons [iBu00]. Depending on the cost, the storage on items like these varies. Often, the wish for a small template is also present due to slow access times of these mediums.

2.3 Robustness

The final requirement which should be mentioned here is that voice authentication is supposed to work even if the conditions are comparatively bad. This applies, for example, to the presence of distinct kinds of background noises in the recordings. Causes for this can be a noisy environment as well as the equipment used to make the

recordings, especially because BioID should be able to work with cheap microphones and off-the-shelf sound cards.

2.4   Accuracy

In a biometric authentication system, there is often a trade-off between user convenience and recognition accuracy in terms of error rates of the system. For example, the requirement of small utterances mentioned above increases the convenience for the users. On the other hand, more speech material in training as well as in test enables a more reliable authentication. Therefore, the goal of speaker recognition in BioID is to implement a classification system that maximizes recognition accuracy, taking constraints that result from convenience requirements and the actual use of the system in a real-world setting into account.

3   The choice of the classification technique

Apart from the feature extraction, which is described in chapter 5, the probably most important decision in the design of a pattern recognition system is the choice of the classification technique. Research in the field of speaker recognition has been done for several decades now, and distinct approaches in regard to the classification have been pursued. In the following, four distinct algorithms are described and compared. Finally, one of these techniques is chosen because it fits best to the requirements that have been mentioned above.

3.1   Vector Quantization (VQ)

Vector quantization, which is also used for speech coding, uses some training set of speech recordings to estimate a *code book*. This contains the means of clusters in feature space. In order to be used for speech coding and compression, cluster mean values are numbered such that they can be identified by indexing. In order to compress a speech signal, each feature vector is assigned to the nearest cluster mean, making it possible to represent this vector by its cluster index only. For reconstruction of an approximation of the original sequence, cluster means are used instead of the original feature vectors. In order to retrieve the signal in time domain, a reversible feature extraction technique has to be used. The quantization error in feature space is the mean distance between the feature vectors computed from the original speech signal and the code book cluster means of the reconstruction [LBG80].

The observation that the quality of speech coding with a code book is highly dependant on the similarity between the training set and the coded material can serve as a motivation for the use of code books for speaker recognition. In this case, for each speaker, a code book is estimated in training. This code book can be thought of as containing those features as mean vectors which are characteristic for that speaker. Classification of unknown signals is based on the mean quantization error of test feature

vectors in regard to the appropriate speaker specific code books, i.e. the quantization error is used as a distance measure.

3.2 Gaussian Mixture Models (GMM)

Another common approach in the field of speaker recognition is to use gaussian mixture models. In general, a well-known approach for the solution of classification problems is the estimation of class specific probability density functions (PDFs) and classification using a-posteriori-probabilities of the presence of classes taking the observed sensor data into account. Under certain assumptions, this is known as the Bayes classification [Nie90].

The most important challenge of this technique in practice is the estimation of probability density functions on the basis of the training data. This is done by choosing a suitable family of PDFs which is able to estimate the "real" PDFs. Which family of functions is suited depends on the application and the kind of feature extraction that is used. In regard to speaker recognition, it has been shown in the literature that gaussian mixture models are well-suited [Rey95].

A gaussian mixture consists of several single gaussians. A multidimensional, single gaussian PDF of feature vectors x depends on the mean vector μ and the covariance matrix Σ and can be written as follows:

$$N(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$$

The gaussian mixture representing class Ω_κ is a weighted sum of n single gaussians:

$$p(x \mid \Omega_\kappa) = \sum_{\nu=1}^{n} c_{\kappa\nu} N(x \mid \mu_{\kappa\nu}, \Sigma_{\kappa\nu})$$

Using GMMs as speaker models can be thought of as a refinement of VQ codebooks, because each single gaussian represents a cluster of feature vectors similar to a codebook cluster which is modeled more accurately.

3.3 Dynamic Time Warping (DTW)

VQ and GMM classification have in common that test feature vectors of a sequence are evaluated in regard to the model without taking the chronology of the vectors into account. Although this is adequate for text independent speaker recognition, which means that the content of spoken utterances in training and test differ, this is not optimal for pass phrase based authentication, as this is a text dependent application.

Dynamic time warping is a classification approach based on distances in feature space, similar to VQ quantization error, but makes use of the fact that in training, the same pass phrase is spoken as in test. DTW compares the test vector sequence to a stored sequence

from training directly, taking into account that two utterances of the same word or phrase are never exactly identical as distinct phonemes can be spoken shorter or longer. In order to cope with that, a time alignment of test and training sequences is found which is optimal in the sense that there is no other alignment yielding a smaller overall distance and fulfilling certain restrictions.

3.4 Hidden Markov Models (HMM)

Finally, the classification using Hidden Markov Models can be thought of as a combination of the text independent GMMs and text dependent DTW. An HMM consists of several states, each state modeling a specific part of an utterance. The distribution in feature space that corresponds to a particular state is modelled statistically, e.g. with a GMM. A disadvantage of this approach is that robust estimation is difficult if very little training data is present. Although this depends on the type and parameterization, e.g. the number of states, a large number of parameters have to be estimated, which can lead to an inexact model.

3.5 Summary and selection of the classification approach

So, which of the four classification approaches is suited best for pass phrase based speaker recognition in BioID? VQ and GMM do not take advantage of the text dependency. Due to the limited length of the speech material, these techniques are not sufficiently accurate for the use in BioID. Having this in mind, the choice between the text dependent techniques DTW and HMM remains. Accurate training of an HMM with as little as five pass phrase utterances seems hardly possible. Therefore, DTW has been chosen in BioID for classification.

4 VQ for template compression

Another requirement which was stated at the beginning is a small biometric template, enabling the storage on, for example, smart cards. When using "standard" DTW for classification, though, this requires to store the full feature vector sequences of all training utterances within the template. With the parameterization used by BioID, this size is approximately 10 kBytes for one single utterance, i.e. about 50 kBytes for the minimal number of training patterns. This is too much for most smart cards on the market today.

In order to solve this problem, the training feature sequences are compressed using the vector quantization approach as described above in chapter 3. All the training utterances of one speaker are used to estimate a code book with 16 cluster means. This code book and the indizes for reconstruction of the training vector sequences are stored in the biometric model of a speaker. With this technique, a compression of the template size for the voice trait to approximately 1 kByte is achieved.

5   MFCC optimized for speaker recognition

As mentioned above, the probably most important parts of any pattern recognition system are the feature extraction and the classification. The goal of the classification is to make a decision based on the test pattern to be recognized and the patterns of known classes from training. The feature computation should make the task of the classifier easier by emphasizing the information in the raw sensor data being most relevant for discriminating classes and discarding the information which is irrelevant in regard to the class. Especially information which does not help to discriminate classes but is different in distinct patterns of the same class should not be present any more after feature processing. To put it another way, inter-class distance should be increased and intra-class distance decreased. Most often, feature extraction leads to a reduction of the dimension of the data, as for most classification techniques, a very high feature dimension makes recognition more difficult [Nie90].

In regard to feature processing of speech signals in general, mel frequency cepstral coefficients (MFCC) are used most often [Schu95]. MFCCs are commonly used for speech recognition purposes as well as for the classification of the speaker, as they contain information about speaker-specific properties of the speech signal as well as about which phonemes have been uttered. For the use in BioID, both of these aspects are in principle valuable, as the speaker recognition is text dependent. Nevertheless, if the classification decision depends too much on the recognition of the text and not on the individual voice characteristic features of a person, this would increase chances of an impostor who has knowledge of a valid pass phrase. Therefore, the parameterization of MFCC computation has been changed from what is commonly used for speech recognition purposes. Feature dimension and frame size have been optimized taking the conditions like the kind of recordings, utterance lengths and the use of DTW into account. A longer framesize makes a higher spectral and cepstral resolution and a higher feature dimension possible. It has been shown experimentally that lower cepstral coefficients tend to contain more information about the phonemes, whereas higher coefficients are more relevant for speaker specific voice characteristics. The figure shows a visualization of a feature vector sequence of a typical pass phrase utterance with this parameterization. Notable are especially the light areas, which are characteristic for a specific phoneme spoken by a particular person.
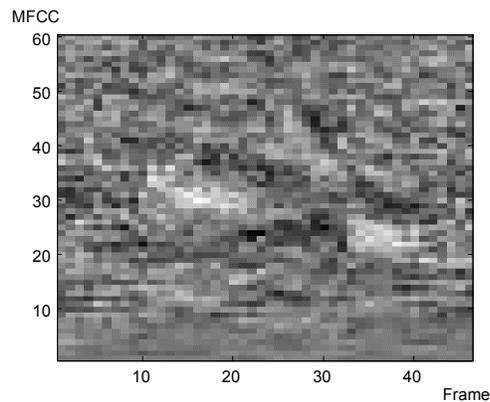

Figure 1: MFCC feature vector sequence visualization of a pass phrase utterance

6   Increasing robustness:
Speech enhancement and cepstral mean subtraction

The last requirement, which has not been mentioned in regard to the description of the feature extraction and classification algorithms is the robustness of the system against difficult conditions. This includes noise or distortions resulting from imperfections of the recording hardware and the use of BioID in situations where ambient noise is present.

6.1   Normalization techniques

Distinct approaches for normalization have been evaluated. The goal of these techniques is to make the data more homogeneous, reducing the influence of unwanted effects. Normalization can be applied at several different points within the classification system:

- Speech enhancement is a pre-processing technique which reduces background noises of the signal before feature extraction. This method takes the original signal in time-domain as input, transforms it to the spectral domain where an estimation of the noise is spectrally subtracted, and finally transforms it back to the time domain.

- Cepstral mean subtraction (CMS) achieves a normalization in feature space. The mean value of all vectors of a sequence is computed. This mean is subtracted from all feature vectors. Therefore, the resulting vector sequence has a mean vector of zero.

- Finally, a normalization of the scores as computed for the test utterance by the classifier can be done.

For these three normalization techniques, experiments have been done to evaluate recognition performance for distinct conditions. In order to gain results which are most relevant for the task of recognizing the speaker characteristics rather than being able to distinguish distinct spoken words, text independent VQ has been used for classification. As better recognition rates are achieved with DTW classification, as it is actually used in BioID, the results are intended for relative comparison rather than as an absolute
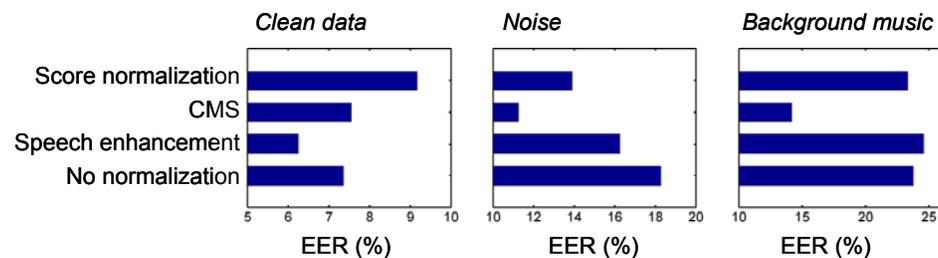

Figure 2: Comparison of normalization techniques under several distinct conditions

measure of system performance.

In order to be able to have comparable results for distinct conditions, a corpus with mostly clean recordings of pass phrase utterances has been used and several kinds of background noises have been superimposed. As a measure for recognition performance, the equal error rate (EER) of a verification decision, which is observable if the system is adjusted such that the false acceptance rate equals the false rejection rate, is used [Mar97]. In the graphics, EERs are compared for recognition without any normalization, with speech enhancement, cepstral mean subtraction and score normalization. This comparison of the three techniques for increasing robustness is provided for the "clean" recordings and for variants of the data base that have been produced by mixing the original waveforms with white noise and music respectively.

6.2   Selection of normalization methods

Experimental results show that there is no one single best normalization technique. Both CMS and speech enhancement achieve lower error rates in two of three conditions. In the third one, only a very slight loss in recognition accuracy can be observed. Score normalization, however, results in a noticeable increase in EER on the clean recordings. Furthermore, a calibration of the score normalization method is necessary which is difficult to achieve in practice. Therefore, for use in BioID a combination of CMS and speech enhancement has been chosen, which is optimal in most cases.

7   Concluding experimental evaluation

Applying the described methods, recognition performance has been evaluated experimentally. There are a number of speech corpora intended for speaker recognition experiments. Some of them are useful for text independent recognition of speakers, as they consist of telephone conversations. There are a few corpora that contain prompted digits and might be used for text dependent speaker recognition, but were recorded with telephone handsets also [Cam99]. No data base and evaluation protocol is known to the authors that is appropriate for the task of pass phrase based speaker recognition under the conditions that are valid for BioID. Therefore, a data base was recorded especially to reflect these conditions like utterance lengths and sampling rate. The corpus includes 22 distinct speakers, each of them uttering their pass phrase ten times. To evaluate the recognition performance of the voice trait as implemented in the SDK, the configuration of pre-processing, feature extraction and classification that is used within BioID has been used for recognitions on the data base. Five recordings of the pass phrase of a particular speaker are used for training, which is the minimal number of utterances that the BioID system requires for enrollment. The remaining utterances are used for test. As a single test utterance is used for multiple verification decisions in conjunction with all the speakers in the data base, a total of about 2,400 verification trials have been evaluated. This results in an equal error rate of 2.7%. As BioID is a multimodal system, the combination with the other biometric traits of course yields an even smaller overall error rate.

It should be kept in mind that the error rate on an appropriate corpus as given above is only one aspect that has to be considered when evaluating the security of a biometric system. For example, the vulnerability against replay attacks is not reflected in the equal error rate of the system. In regard to BioID authentication, it might be theoretically possible to gain access with a recorded utterance of a legal user known to the system. However, the multimodality of the system enhances security in this respect again, as it would be necessary to record not only an utterance of the user's pass phrase, but a video image of the trained person while uttering the phrase as well. Therefore, an intrusion on the basis of replaying a recording seems quite unlikely, even with a considerable amount of effort on the side of an attacker.

8   Summary

Although most of the techniques described in this article have been known from the literature for quite some time, the challenge of speaker recognition lies in the combination and optimization of these techniques for use under real-world conditions. Only if it is possible to use a biometric authentication system with little effort, if the accuracy is sufficiently high and if it is robust against imperfect sensor data, biometric technology can leave the scientific laboratory and let users take advantage of the principal benefits biometric authentication has over token-based or knowledge-based methods of authentication.

Pass phrase based speaker recognition in BioID uses dynamic time warping for text dependent speaker recognition. A verification or identification is achieved with as little as a single one second test utterance. Only five seconds are sufficient for training. Better recognition accuracy even with limited speech material is achieved by optimizing parameterization of MFCC features, higher dimension and longer frames leading to increased cepstral resolution. Robustness is gained by a combination of speech enhancement in spectral domain and cepstral mean subtraction, techniques which complement each other and decrease the influence of distinct kinds of interferences of the speech data. Finally, a small biometric template size, which is important for the use of BioID with smart cards and similar media, is enabled because vector quantization with speaker specific codebooks is applied for model compression. On a data base which is appropriate for the kind of utterances and recording conditions, a verification equal error rate of 2.7% has been achieved.

9   Bibliography

[Bio00]   Frischholz, R.; Dieckmann, U.: „BioID: A Multimodal Biometric Identification System“, IEEE Computer, Vol. 33, No. 2, February 2000.
[Cam99] Campbell, J. P. Jr.; Reynolds, D.A.: "Corpora for the Evaluation of Speaker Recognition Systems", in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 1999.
[iBu00]   "DS1955B Java-powered Cryptographic iButton", Dallas Semiconductor Corporation, August 2000, http://www.ibutton.com/software/crypto/fips140-1l3.pdf

[LBG80] Linde, Y.; Buzo, A.; Gray, R.: „An Algorithm for Vector Quantizer Design", in IEEE Transactions on Communications, Vol. 28, Nr. 1, January 1980, pp. 84 – 95.

[Mar97] Martin, A.; Doddington, T. K. G.; Ordowski, M.; Przybocki, M.: "The DET curve in assessment of detection task performance", in Proceedings of EuroSpeech '97, volume 4, pp. 1895 – 1898, 1997.

[Nie90] Niemann, H.: „Pattern Analysis and Understanding", Second Edition, Springer Series in Information Sciences 4, Springer Verlag, Heidelberg, 1990.

[Rey95] Reynolds, D.; Rose, R.: "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", in IEEE Transactions on Speech and Audio Processing, Vol. 3, No.1, pp. 72 – 83, January 1995.

[Schu95] Schukat-Talamazzini, E. G.: „Automatische Spracherkennung", Vieweg Verlag, March 1995.