Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition 39 (2006) 147–155

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and brief communication

Combining classifier decisions for robust speaker identification

Daniel J. Mashao*, Marshalleno Skosan

Speech Technology And Research (STAR), University of Cape Town, Rondebosch 7701, South Africa

Received 23 June 2005; received in revised form 17 August 2005

Abstract

In this work, we combine the decisions of two classifiers as an alternative means of improving the performance of a speaker recognition system in adverse environments. The difference between these classifiers is in their feature-sets. One system is based on the popular mel-frequency cepstral coefficients (MFCC) and the other on the new parametric feature-sets (PFS) algorithm. The feature-vectors both have mel-scale spectral warping and are computed in the cepstral domain but the feature-sets differs in the use of spectral filters and compressions. The performance of the classifier is not much different in recognition rates terms but they are complementary. This shows that there is information that is not captured in the popular mel-frequency cepstral coefficients (MFCC), and the parametric feature-sets (PFS) is able to add further information for improved performance. Several ways of combining these classifiers gives significant improvements in a speaker identification task using a very large telephone degraded NTIMIT database.

Keywords: Speaker identification; Parametric feature sets; Multiple classifier systems; Gaussian mixture model

1. Introduction

Over the years, there has been much interest in using speech as a means of identifying speakers. Speech, as opposed to other biometrics such as fingerprints and face recognition, allows recognition to be performed remotely as it can easily be transmitted over communication channels. It has been shown, however, that the performance of speaker recognition (SR) systems degrades considerably when contaminated by telephone noise in transmission [1]. Hence, the robustness of SR systems has been a major research issue in recent years [2]. For SR tasks, numerous speech parameterisations and classification algorithms have been proposed over the years [3]. However, it is still difficult to implement a single classifier that exhibits sufficiently high performance in practical applications. As a result, several researchers have cited the fusion of multiple information sources as a promising option in SR research [3–5]. A sufficient condition for fusing the outputs of many classifiers is that they make errors that are uncorrelated i.e. they misclassify different patterns of the same data [6]. Altincay and Demirekler [7], have improved speaker identification (SI) performance by fusing the outputs of two classifiers where one used a form of channel compensation and the other did not. They showed that SI performance is very sensitive to the signal processing done when extracting a particular speech features. In this work we develop two high performing baseline SI classifiers that make different errors. Their decisions are then subsequently combined with the aim of improving the robustness and performance of the overall SI system. In particular, large population speaker identification experiments are conducted on the telephone degraded NTIMIT speech database. Large improvements, above the baseline SI classifiers and other systems reported in related literature, are obtained.

2. Features from the speech signals

A speaker identification systems consists of mainly two parts as shown in Fig. 1. In the front-end is the feature-generation part and in the back-end is the classification engine. During the enrollment phase the switch is turned

* Corresponding author. Tel.: 27 216502816.
E-mail addresses: daniel@star.za.net (D.J. Mashao), leno@star.za.net (M. Skosan).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.08.004

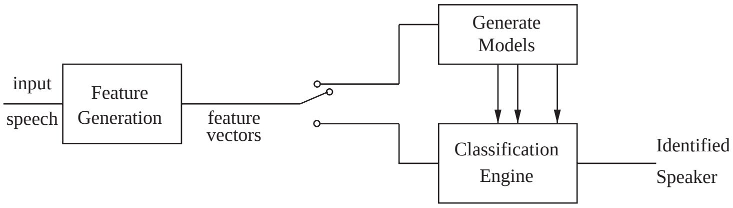

Fig. 1. Speaker identification system.

towards the upper route and the system generates models, and during testing or evaluation phase the models are used to identify a speaker from an unknown speaker's speech. Current state of the art systems uses the mel-frequency cepstral coefficients (MFCC) algorithm for the front end and Gaussian mixture models (GMM) and the classification engine.

An ideal front-end in the SI system is the one that will extract all speaker specific information from the input speech signal without being confused with what the person is saying. Every person has a natural sound quality due to their voice pitch. However, pitch detection has proven challenging and also reliance on it can allow impostors to gain access by changing their own pitch. The other problem with pitch is that it cannot be reliably measured in some speech sounds for example nasals and consonants. As such many front-end algorithms do not seek to use pitch as a specific feature. Instead the design of speech feature-sets seeks to find speaker specific information in other parts of a person's speech.

2.1. Mel-frequency Cepstral Coefficients

The MFCC is the most popular front-end for SI systems and is also used in other speech technology tasks such as speech recognition and SR in general. The MFCC coefficients are generated as shown in Fig. 2. First, the speech signal is acquired in the time domain via sampling and after the application of the discrete Fourier transform (DFT) it is converted into the frequency domain. If the inverse discrete Fourier transform (IDFT) was applied then, the signal would revert back to the time domain, but before the IDFT is applied a log magnitude (taking the logarithms of the magnitude of a complex signal) of the frequency domain signal (the spectrum) is computed. The signal is now said to be in the cepstral (a play on the words spectral) domain and the units in this domain are seconds, same as in the time domain. The signal in the cepstral domain is measured in quefrencies (again a play on the words frequency). The low-order quefrencies contain information that is due to the speech formants and therefore carries information about what is being said and the high quefrencies are due to the pitch and therefore assumed to be speaker dependent.

To obtain the MFCC coefficients, the speech signal is windowed and converted into the frequency domain by using the DFT. In the frequency domain a log magnitude of the complex signal is obtained. A mel-scaling or mel-warping is then performed using filters. The common method of implementing these filters is to use triangular filters that are linear spaced from 0 to 1 kHz and then non-linearly placed according to the mel-scaling approximations. There are several approximations of the mel-scale the most popular is by O'Shaughnessy [8] which is

$$F_{mel} = 2595 \log \left(1 + \frac{F_{in}}{700}\right),$$

where F_{mel} is the frequency in mels and F_{in} is the input frequency in Hertz. This is the scaling used in the design of MFCC coefficients used in this paper. There are many functional approximations of the mel-scale and they all show minor differences in performance as shown by Umesh et al. [9]. The centre frequencies of the triangular filters will be set at the mel-scale frequencies, with the low input frequencies (less than 1 kHz) given a higher profile than the higher frequencies. The resultant signal from the filtering is then transformed using an inverse DFT (usually implemented with a discrete cosine transform) into the cepstral domain. The lower order coefficients are selected as the feature vector. The selection of the lower order coefficients is done on purpose to avoid the higher coefficients which include the pitch. The coefficients are then uniformly scaled and used as the output feature vector for that speech frame.

It is worth noting that these same MFCC feature-vectors are used in both speech recognition and speaker recognition tasks. In speaker independent speech recognition task any speaker information is considered noise but in SR it is actually the kind of information that is sought. The fact that the same feature-vectors can be used for both tasks shows that they contain both semantic and person specific information. The successful application of the low order quefrencies coefficients for SI tasks shows that the person information is still intact.

The MFCC has been used for several years since the late 1990s and has successfully replaced the linear prediction cepstral coefficients (LPCC). The main advantage of the LPCC was computation but with increasing computing power and better performance (and the use of the fast Fourier transform—FFT), the MFCC has completely dominated the designs of the front-ends of speech technology systems. Competition from the auditory based feature-sets has not been successful and mainly due to their lower performance and very high computational cost. Our own research

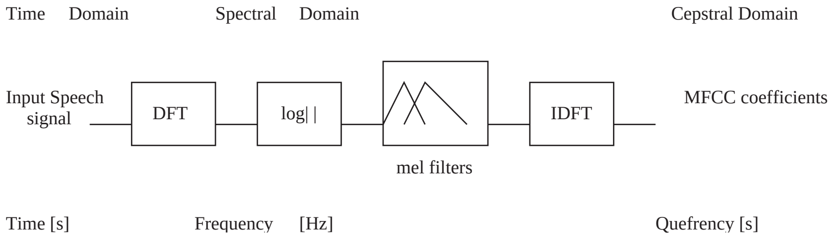

Fig. 2. Mel-frequency cepstral coefficients.

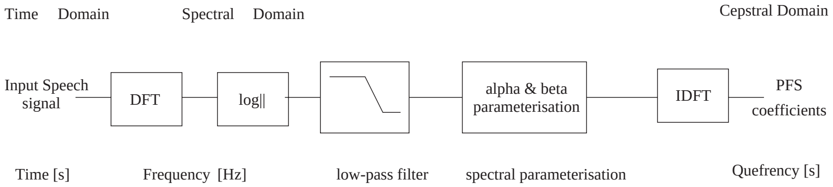

Fig. 3. PFS feature vector generation.

found that the ensemble interval histogram (EIH) method under performed the MFCC but only in cases of data mismatch were the performance is bad anyway was there merit for the EIH algorithm. The same conclusion was reached by the authors of the EIH [10,11].

2.2. Parametric feature-set

An alternative method that is similar to MFCC is the parametric feature-set (PFS). The PFS works similar to MFCC but does other parts differently and will be discussed in this section. The PFS is computed as shown in Fig. 3. As shown in the figure the first part is similar between the MFCC and PFS. The PFS applies the low-pass lifter (play on words filter) and applies a parameterisation which is almost similar to controlling the centre frequencies of the MFCC triangular filters. The PFS method can be stated as follows:

(1) Hamming window a frame, convert into the frequency domain via DFT.
(2) Compute the log magnitude.
(3) Apply a low pass filter.
(4) Select the samples for a new signal using the parameterisations (α, β). These parameters (α, β) will be explained in detail later.
(5) Convert the new signal into the cepstral domain.

All the basic ideas of MFCC are preserved in the PFS. For example the feature-vector is computed in the cepstral domain, the use of the FFT and the selection of low order quefrencies. The PFS, however, explicitly removes the pitch information by using a low pass lifter in the frequency domain. The pitch information is removed before conversion into the cepstral domain and this is a major difference between the MFCC and PFS, because even if the parameterisation (α, β) was not done, the pitch information will not be present in the cepstral domain. In the case of MFCC this information is there but not included in the feature vector by transforming into the cepstral domain of a low order or by selecting only low order cepstral coefficients.

The main advantage of the PFS is that it simply allows the designer to control the amount of spectral warping applied to the speech spectrum. This is done by adjusting two parameters α and β according to the equation,

$$A\sum_{i=1}^{\alpha}\beta^{i-1}=\frac{N}{2},$$

where N is the length of the speech spectrum. The α term is the number of regions in which the $N/2$ spectrum is to be divided. For example when $\alpha=1$ it means that the spectrum is not to be divided at all. No spectral warping will be achieved in that case. For $\alpha=2$ it means the spectrum is to be divided into two regions of size A and $A\beta$, for $\alpha=3$ means the spectrum is to be divided into three regions of size A, $A\beta$ and $A\beta^2$, for $\alpha=4$, the spectrum is to be divided into four regions, A, $A\beta$, $A\beta^2$, and $A\beta^3$, and so on for various values of α. All these regions are to receive the same number of samples equally spaced in the region. It is therefore, clear that like in the case of MFCC the lower frequencies of the spectrum will be emphasised at the expense of the higher frequencies since the lower frequencies will receive more 'sampling' in their region. In effect the PFS is achieving a more flexible version of spectral warping. For more detailed information see [12]. In brief, the parameters α and β control the amount of spectral warping applied to the speech spectrum. For example, if $\alpha=4$ and $\beta=2.0$, the amount of spectral compression exhibited by the parameterised feature set approximates that of the mel-scale compression or spectral warping. For $\beta=1$ there is no spectral warping irrespective of the value of α. In this work, we use $\alpha=4$, $\beta=1.7$, since it

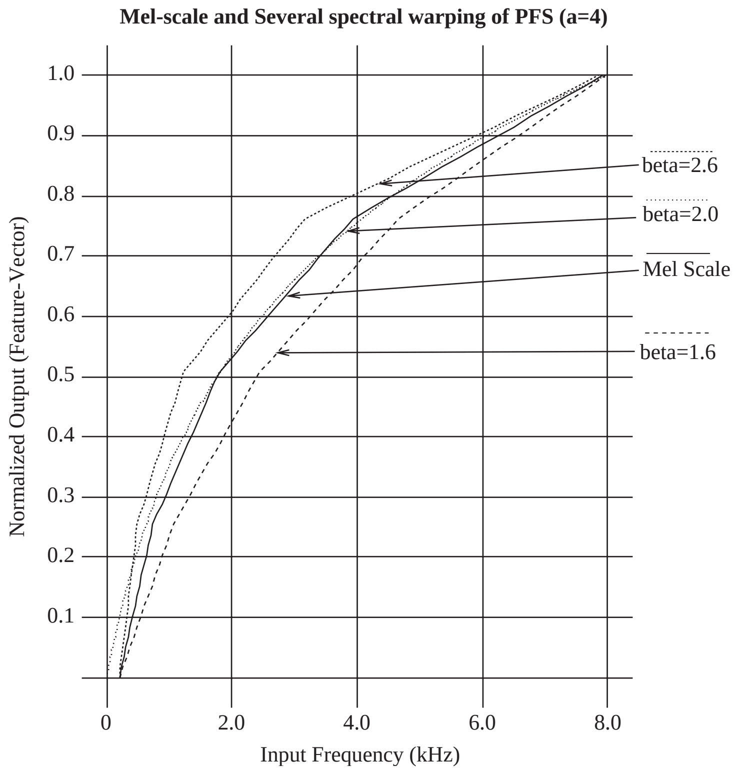

Fig. 4. Comparison of spectral warping of different PFS ($\alpha = 4, \beta$) and the mel-scale.

was shown to produce the best SI performance [13] on the NTIMIT database. Fig. 4 shows the relationships caused by various β values on the spectral warping. The other common parameters between MFCC and PFS is framing of the input speech signal, the sampled speech signal is multiplied by overlapping Hamming windows which divided it into a sequence of 20 ms frames with a step size of 10 ms between frames.

Therefore, in summary for the case of MFCC, the speech frames are transformed into the frequency domain where a sequence of log-magnitude spectra are computed. The mel-scale warping is obtained by using mel-scaled triangular filter banks. The outputs of the filter-banks are then transformed into the cepstral domain and the 30 low-order coefficients are scaled to an appropriate range (by multiplying with constants e.g. 2000) and form the feature vectors. For the parameterised feature set, the log-magnitude spectra are first smoothed with a low-pass FIR filter. The filtered log-magnitude spectra are then non-linearly sampled according to parameterisation equation and the resultant signal is converted into the cepstral domain using a discrete cosine transform. Similarly to the MFCC, only the first 30 low-order coefficients that are appropriately scaled are used as the feature-vector for that frame. For both feature extraction algorithms, the first component of each feature vector is discarded and time derivative parameters are not used as the addition of these were found to reduce performance. In order to model the distribution of feature vectors obtained for each speaker, we used GMM [2,14], and it is described next.

3. Gaussian mixture models

A GMM can be viewed as a non-parametric, multivariate probability distribution model that is capable of modelling arbitrary distributions and is currently the dominant method of modelling speakers in SR systems. The GMM of the distribution of feature vectors for speaker s is a weighted linear combination of $M = 32$ unimodal Gaussian densities $b_i^s(\vec{x})$, each parameterised by a mean vectors μ_i^s each of dimension $D = 30$ with a diagonal covariance matrix Σ_i^s. These parameters are collectively represented by the notation

$$\lambda_s = \left\{ p_i^s, \mu_i^s, \Sigma_i^s \right\}$$

for $i = 1, 2, \ldots, M$, and are known as a speaker's model. The p_i^s are the mixture weights satisfying the constraint

$$\sum_{i=1}^{M} p_i^s = 1.$$

For a feature vector $\vec{x}$ the mixture density for speaker s is computed as

$$p\left(\vec{x}|\lambda_s\right) = \sum_{i=1}^{M} p_i^s b_i^s\left(\vec{x}\right),$$

where

$$b_i^s(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i^s|^{1/2}} \times \exp\left(\frac{1}{2}(\vec{x} - \mu_i^s)'(\Sigma_i^s)^{-1}(\vec{x} - \mu_i^s)\right)$$

and $D = 30$ is the dimension of the feature-space.

Given a sequence of feature vectors $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, for an utterance with T frames, which for the purpose of computations are assumed to be independent, the log-likelihood of a speaker model s is

$$L_s(X) = \log p\,(X|\lambda_s) = \sum_{t=1}^{T} \log\ p(\vec{x}_t|\lambda_s).$$

For speaker identification, the value of $L_s(X)$ is computed for all speaker models λ_s enrolled in the system and the owner of the model that generates the highest value is returned as the identified speaker. In this work, we modelled speakers by using GMMs with 32 mixtures (a Gaussian is referred to as a mixture) according to the work done by Reynolds in [1].

4. Experimental set-up

The NTIMIT database [15,16] was used for all the experiments conducted in this research. This database consists of 630 speakers (438 males and 192 females). Each speaker is recorded while reading 10 phonetically balanced sentences of roughly 3 s each. The NTIMIT database is a network or more appropriately a telephone noise version of the TIMIT database [17]. The TIMIT speech was then played through a carbon button telephone handset and transmitted over local and long distance telephone channels [1] to generate the NTIMIT database. Even though the database was generated on the 4 kHz bandwidth telephone channel it is still sampled at 16 kHz like the original TIMIT database. According to the Language Data Consortium [18] (LDC), the TIMIT database is the most requested speech database indicating its popularity in speech technology research. The NTIMIT database ranks fourth in the list of the top ten requested speech databases according to the LDC information.

The TIMIT database is collected from American English speakers divided into eight (8) accent regions including speakers that do not have strong regional accents. The database is divided into train and test sets by its creators to enable comparisons of results from different researchers. The train and test set is usually used for speech recognition tasks. For the text independent SI tasks it is much easier to use the whole database rather than to follow the division of test and train sets. Each speaker said 10 different utterances. These 10 utterances that are read can be divided into three groups. The first two utterances are common across all speakers in the database and are known as the sa1 and sa2 utterances. These utterances can be used for speaker normalisation and accent identification. The next eight utterances are different across all the speakers and are known as the sx and si utterances. The common way of using this database is to use the first eight utterances (including the sa1 and sa2 utterances) of each speaker for the model training and the last two utterances for SI evaluation. This is how the NTIMIT database is used in all the results reported in this paper in the design of a text independent speaker identification system for both classifiers.

Table 1
Speaker identification rates on NTIMIT databases

System	Speaker identification rate (%)
Reynolds [1]	60.7
Le Floch et al. [19]	58.0
Mashao and Baloyi [13]	69.2
Lerato [20]	71.1

Table 2
Baseline classifier performances

Baseline classifier	Speaker identification rate (%)
Type of front-end	
MFCC	71.6
PFS ($\alpha = 4.0$, $\beta = 1.7$)	70.6

General performance on this database varies between a speaker's identification rate of 58% and 72% as indicated by the Table 1. The weakness of the TIMIT database is that all the recordings were done in a single session. This is not much of a problem for the reported results as the aim is to show how the combination of very similar classifiers that differs in how and when is the high quefrencies information is eliminated.

The performance of the baseline systems based on MFCC and PFS are shown in Table 2.

5. Multiple classifier systems

A substantial amount of empirical evidence has shown that multiple classifier systems can be used to enhance a number of pattern recognition applications [3,6,7,21,22]. Doddington et al. [3] reported that in the NIST 1998 SR evaluations a number of the participants improved baseline system performance by a simple linear combination of the scores obtained for different system. Chen and Chi [23] applied a novel method of combining multiple probabilistic classifiers using different feature sets extracted from the same raw speech signal to a speaker identification task. They showed that the combination classifiers based on different spectrum representations can be used to improve the robustness of SI systems. However, they performed SI experiments on clean speech for a population of only 20 male speakers. Ramachandran et al. [24] provided a discussion of how various forms of diversity, redundancy and fusion can be used to improve the performance of SR systems. Their experiments showed improvements in speaker verification performance when forming a simple linear combination of three different classifiers based on the same front-end features. SR is an area of pattern recognition that is well suited to the application of multiple classifier systems as there exists a number of feature extraction techniques and classification algorithms, each with its own advantages and disadvantages.

Consider for example the channel compensation method known as cepstral mean normalisation (CMN). It has been shown to improve performance when evaluated on speech distorted by convolutional channel noise. However, experiments on clean speech show a reduction in SR performance [25]. This suggests that although CMN can be used to improve SR performance on telephone speech it does remove some speaker specific information as well. Experiments by Altincay and Demirekler [7] show that we can take advantage of the strengths of this method, while compensating for its weaknesses, by fusing two replicas of the same speaker identification system, with and without the application of CMN. There are many ways of constructing multiple classifier systems. These include using different training sets, different feature sets, different classification algorithms and different classifier architectures and parameters [21].

Multiple classifier systems are generally classified according to their architecture, the type of outputs produced by their constituent classifiers and the techniques used to combine them. The architecture of a multiple classifier system can be parallel, serial or an arbitrary combination of the two (hybrid architecture). In a parallel architecture all the base classifiers are invoked independently and a single combination function is used to combine their outputs. In a serial architecture all the base classifiers are invoked in a linear sequence with each successive classifier producing a reduced set of possible pattern classes [26]. The base classifiers in multiple classifier systems usually produce outputs at one of three levels of information namely, the abstract level, the rank level and the measurement level [22]. At the abstract level, each base classifier outputs a unique class label for the given input pattern. At the rank level, each base classifier outputs an ordered list of possible classes for each input pattern. The class at the top of the list is said to have the highest rank. At the measurement level, each base classifier outputs a numerical value for each class as an indication of the confidence that it has that the input pattern belongs to that class. The methods used to combine the various levels of base classifier output generally fall into two categories namely, fixed rules and trained rules. Fixed rules are static in that their form and parameters do not change as a result of the output produced by the base classifiers. As such, they are simple, have low time and memory requirements and are well suited to groups of classifiers that exhibit similar performances and make uncorrelated errors. On the other hand, trained rules adapt their parameters to the outputs of the base classifiers and as such, at times, perform better than fixed rules. They are said to be more suitable than fixed rules for combining classifiers that have different levels of performance and make correlated errors. These rules generally have high memory and time requirements and make heavy demands on the quality and size of the training set. For more detailed information on multiple classifier systems, the interested reader is referred to Rolis tutorial on the subject [26].

As stated earlier, many empirical studies have shown that multiple classifier systems can improve the performance of a range of pattern recognition applications. Unfortunately, there are fewer studies that provide a sound theoretical underpinning for the improvements gained in multiple classifier systems. One such study however, was done by Kittler et al. [6] in 1998. They provided a common theoretical framework for combining classifiers which use distinct pattern representations to estimate the posterior probabilities of the given input patterns. Under the assumption that the representations used are conditionally statistically independent and, that the posterior class probabilities computed by the base classifiers do not differ greatly from the prior probabilities, a number of fixed combination rules were derived from Bayesian theory. In particular, if we assume that $X^{(1)}$ refers to the feature vectors associated with the base classifier that uses the MFCC front-end and $X^{(2)}$ are the corresponding feature vectors associated the base classifier that uses the PFS front-end, we can apply what we abbreviated as the KHDM rules after the authors Kittler, Hatef, Duin and Mataz [6] to speaker identification as follows:

The sum rule

$$\hat{L}_{sum}(X) = \arg \max_{s=1}^{N} \left[\sum_{i=1}^{2} L_s(X^{(i)}) \right],$$

the product rule

$$\hat{L}_{prod}(X) = \arg \max_{s=1}^{N} \left[\prod_{i=1}^{2} L_s(X^{(i)}) \right],$$

the maximum rule

$$\hat{L}_{max}(X) = \arg \max_{s=1}^{N} \left[\max_{i=1}^{2} |L_s(X^{(i)})| \right],$$

and finally the minimum rule

$$\hat{L}_{min}(X) = \arg \max_{s=1}^{N} \left[\min_{i=1}^{2} |L_s(X^{(i)})| \right],$$

where N refers to the number of speakers enrolled in the system. These rules are used to measure performance of the combined classifiers based on MFCC and PFS.

6. Results

In this work, we argue that the signal processing differences between MFCC and parameterised feature sets cause

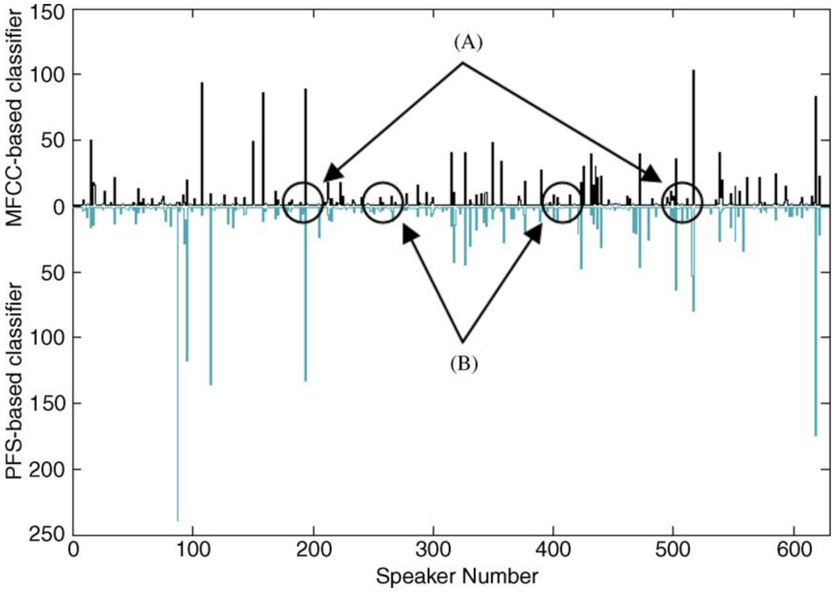

Fig. 5. Rank of the speakers in the two classifiers.

them to extract different information from the speech spectrum and, as a result cause the base classifiers, based on them, to misclassify different speakers. Fig. 5 shows the rank of the true (correct) speaker as decided by the MFCC-based SI system versus the rank of the true speaker as decided by the PFS-based SI system. By rank we mean that if the features extracted from the input speech signal were compared to each speaker model and the base classifier were to output an ordered list of speakers from most likely to least likely, then the position of the true speaker in this list would determine his/her rank. The diagram also shows that both SI systems generally place the correct speaker amongst the most likely speakers. The fact that the correct speaker is usually very high in rank has been noted by other researchers using a GMM-based classification system. This diagram clearly illustrates the difference in the decision profile of the two base classifiers and as such, also aids in showing that the two classifiers make different decisions. Note that for the speakers in the area labelled as (A) in the Fig. 5 it is clear that no combination can recover the error for that particular speaker as both SI systems perform extremely poorly. However, for points labelled as (B) a proper combination of these decisions could reverse the error. After establishing that the two SI classifiers make different errors, we used the KHDM rules to combine them. We used a parallel multiple classifier architecture with measurement level base classifier outputs as input to the KHDM fixed combination rules.

The results in Table 3 show that for the SI base classifiers all the combination rules, except the maximum rule, improves the identification rate above that of the individual base classifiers, with the sum and product rules producing the best results. In their research, Kittler et al. [6] showed that the sum rule outperformed the other combination strategies since it is less sensitive to estimation errors. This could be the reason for the good overall performance exhibited by the sum combination rule. If one uses KHDM rules as is, each classifier is deemed to be equally reliable. However, when we formed a simple weighted linear combination of the baseline SI classifier outputs, a further improvement in the results was noted. In order to find the optimal weights for the linear combination, we performed a simple search of the best weightings. The results of the weighted linear combination are illustrated in Table 4. The weights used to combine the two systems are indicated in brackets for the MFCC and PFS base classifiers respectively and suggests that the MFCC-based classifier is more reliable than the PFS-based one.

From these results it is clear that multiple classifier systems can be used to improve the robustness of SR systems. These results confirm that the two feature extraction

Table 3
Applying KHDM rules on the classifiers

Combination rule	Speaker identification rate (%)
Sum rule	77.0
Product rule	77.0
Minimum rule	72.5
Maximum rule	70.6

Table 4
Using a weighted sum rule

Combination rule	Speaker identification rate
Weighted sum rule	79.1% (0.67, 0.33)

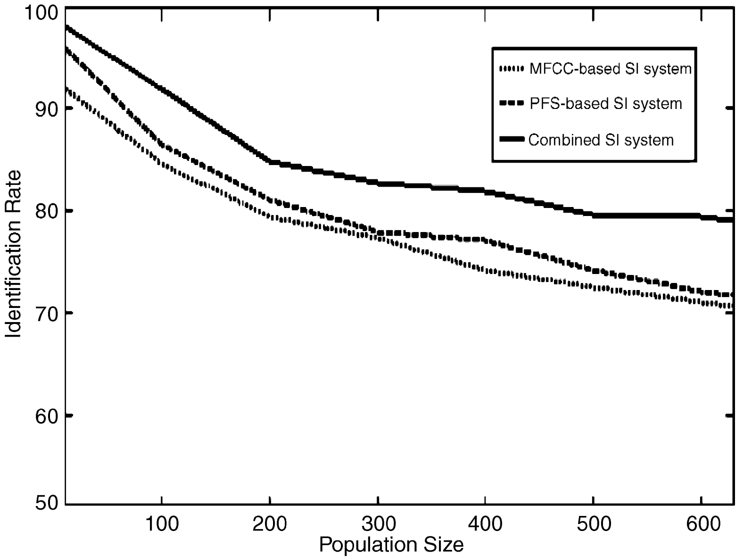

Fig. 6. Speaker identification rate as a function of set size.

algorithms namely, MFCC and PFS extract different information from the speech signal and hence can be fused so as to obtain a more robust speaker recognition system. This also supports Altincay and Demirekler [7] observation that SI systems are very sensitive to the signal processing done in the front-end. The 79.1% identification rate obtained is the highest reported identification rate for the full 630 speaker NTIMIT database with the first eight alphabetically and numerically numbered utterances (including the common utterances) used for training the speakers model and the last two utterances which are different from the ones used in the training are used for evaluation of the speaker identification. Fig. 6 clearly depicts the large improvement in performance obtained when linearly combining the baseline SI classifier outputs using the weights given in Table 4. It shows the speaker identification rate as the population size increases.

7. Conclusions

In this work, we have shown that the combination of multiple speaker recognition systems is an alternative method of improving the robustness and performance of the overall speaker recognition system. The systems only differed in terms of the front-end algorithm even though the algorithms were mainly similar in many ways (their use of the FFT, log magnitude spectra, feature-vectors computed in the cepstral domain) and differed on how the process of computing the spectral warping and dealing with pitch was done. We conclude that the popular MFCC feature-sets is missing some information that can be used to improve their performance. The combined classifiers produced one of the highest reported speaker identification rate performance on the NTIMIT database.

References

[1] D. Reynolds, Large population speaker identification using clean and telephone speech, IEEE Signal Process. Lett. 2 (1995) 46–48.

[2] D. Reynolds, R. Rose, Robust text-independent speaker identification using gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1995) 72–83.

[3] G. Doddington, M. Przybocki, A. Martin, D. Reynolds, The Nist speaker recognition evaluation overview, methodology, systems, results, perspective, Speech Commun. (2000) 225–254.

[4] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian speaker mixture models, Digital Signal Process., DSP, 2000.

[5] J. Campbell, D. Reynolds, R. Dunn, Fusing high- and low-level features for speaker recognition, Eurospeech ISCA, September 1–4, Geneva, Switzerland, 2003, pp. 2665–2668.

[6] J. Kittler, M. Hatef, R. Duin, J. Mataz, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239.

[7] H. Altincay, M. Demirekler, Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence, Speech Commun. 41 (2003) 531–547.

[8] D. O'Shaughnessy, Speech Communication—Human and Machine, Addison-Wesley, New York, 1987.

[9] S. Umesh, L. Cohen, D. Nelson, Fitting the mel scale, in: ICASSP-99, vol. 1, 1999, pp. 217–220.

[10] O. Ghitza, Auditory nerve representation as a front-end for speech recognition in a noisy environment, Computer Speech Lang. 1 (1986) 109–130.

[11] S. Sandhu, O. Ghitza, A comparative study of mel cepstra and eih for phone classification under adverse conditions, in: ICASSP 95, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, IEEE, Silver Spring, MD, 1995, pp. 409–412.

[12] D. Mashao, Computations and evaluations of an optimal feature set for an HMM-based recognizer, Ph.D. Thesis, Brown University, Providence, RI, USA, May 1996.

[13] D. Mashao, N. Baloyi, Improvements in the speaker identification rate using feature sets on a large population database, Proceedings of Eurospeech, vol. 4, 2001, pp. 2833–2836.

[14] D. Reynolds, Automatic speaker recognition using gaussian mixture speaker models, MIT Lincoln Lab. J. 8 (2) (1995) 173–192.

[15] J. Campbell, D. Reynolds, Corpora for the evaluation of speaker recognition systems, Proceedings of IEEE ICASSP-99, 1999, pp. 2247–2250.

[16] S.B.C. Jankowski, A. Kalyanswamy, J. Spitz, Ntimit: a phonetically balanced, continuous speech, telephone bandwidth speech database, in: Proceedings of the IEEE International Conference on Acoustic Speech and Signal Processing, vol. 1, IEEE, Silver Spring, MD, 1990, pp. 109–112.

[17] J. Garofolo, Getting started with the darpa timit cd-rom: an acoustic phonetic continuous speech database, in: National Institute of Standards and Technology (NIST), NIST, 1988.

[18] “Ldc: Linguistic data consortium www.ldc.upenn.edu.” last access, October 2004.

[19] J.L. Floch, C. Montacie, M. Caraty, Speaker recognition experiments on the ntimit database, Proceedings of EUROSPEECH 95, Madrid, Spain, vol. 1, September 1995, pp. 379–382.

[20] L. Lerato, Hierarchical methods for large population speaker identification using telephone speech, Master’s Thesis, University of Cape Town, Cape Town, South Africa, 2003.

[21] A. Jain, R. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4–37.

[22] L. Xu, A. Krzyzak, C. Suen, Methods for combining multiple classifiers and their applications to handwriting recognition, IEEE Trans. Systems, Man, Cybern. 22 (1992) 418–435.

[23] K. Chen, H. Chi, A method of combining multiple probabilistic classifiers through soft competition on different feature sets, Neurocomputing 20 (1998) 227–252.

[24] R. Ramachandran, K. Farrell, R. Mammone, Speaker recognition general classifier approaches and data fusion methods, Pattern Recog. (2002), 2801–2821.

[25] D. Reynolds, Speaker identification and verification using gaussian mixture speaker models, Speech Commun. 17 (1–2) (1995) 91–108.

[26] F. Roli, Fusion of multiple pattern classifiers, Eighth National Conference of the Italian Association on Artificial Intelligence, Pisa, Italy, September 2003.