Speaker Recognition Using MPEG-7 Descriptors

Hyoung-Gook Kim, Edgar Berdahl, Nicolas Moreau, Thomas Sikora

Department of Communication Systems
Technical University of Berlin, Germany
[kim, berdahl, moreau, sikora]@nue.tu-berlin.de

Abstract

Our purpose is to evaluate the efficiency of MPEG-7 audio descriptors for speaker recognition. The upcoming MPEG-7 standard provides audio feature descriptors, which are useful for many applications. One example application is a speaker recognition system, in which reduced-dimension log-spectral features based on MPEG-7 descriptors are used to train hidden Markov models for individual speakers. The feature extraction based on MPEG-7 descriptors consists of three main stages: Normalized Audio Spectrum Envelope (NASE), Principal Component Analysis (PCA) and Independent Component Analysis (ICA). An experimental study is presented where the speaker recognition rates are compared for different feature extraction methods. Using ICA, we achieved better results than NASE and PCA in a speaker recognition system.

1. Introduction

A typical speech or speaker recognition system consists of three main modules: feature extraction, pattern classification and decoder with speech modeling. Because feature extraction influences the recognition rate greatly, it is important in any pattern classification task. Feature extraction unifies the features of the same pronunciations by removing irrelevant information and distinguishes between the features of different pronunciations by highlighting relevant information. Among the most widely used features for speaker recognition is a technique based on a short-term spectrum of speech, where Fourier basis speech signals are decomposed into a superposition of a finite number of sinusoids, which are used for speaker recognition. Using such features, it is not always possible to express the domain's statistical structure, but it assumes that all signals are infinitely stationary and that the probabilities of the basis functions are all equal. In contrast, Principal Component Analysis (PCA) [1] and the more recently developed Independent Component Analysis (ICA) [2][3] perform high dimension multivariate statistical analysis. PCA decorrelates the second order moments corresponding to low frequency properties and extracts orthogonal principal components of variations. ICA, on the other hand, is a linear but not necessarily orthogonal transform, which makes unknown linear mixtures of multi-dimensional random variables as statistically independent as possible. It not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. It extracts independent components even if their magnitudes are small, whereas PCA extracts only components with the largest magnitudes. Thus, ICA representation seems to capture the essential structure of the data in feature extraction and signal separation. The spectrum basis generated by techniques such as PCA and ICA and it's projection have been suggested for feature extraction by the MPEG-7 audio group. MPEG-7 [4] is a standardization initiative of the Motion Pictures Expert Group that, instead of focusing on coding such as MPEG-1, MPEG-2 and MPEG-4, is meant to be a standardization of the way to describe multimedia content. Although MPEG-7 focuses on indexing, searching, and retrieval of audio, the low-level feature extraction audio descriptors have very general applicability in describing not only environmental sounds, but also in describing speech.

In this paper, we evaluate a basis projection method using MPEG-7 descriptors for the analysis of speaker variability and for the extraction of low-dimensional speech features.

2. Extracting speech features with a spectrum basis projection of MPEG-7 descriptors

It is widely known that direct spectrum-based features are generally incompatible with classification applications due to their high dimensionality and their inconsistency. To address the problems of dimensionality and redundancy, whilst keeping the benefits of complete spectral representations, MPEG-7 sound recognition frameworks [4][5][6] use a method of projection onto a low-dimensional subspace via reduced-rank spectral basis functions. The system diagram in Figure 1 shows the MPEG-7 extraction scheme for speech spectrum basis and speech recognition features.

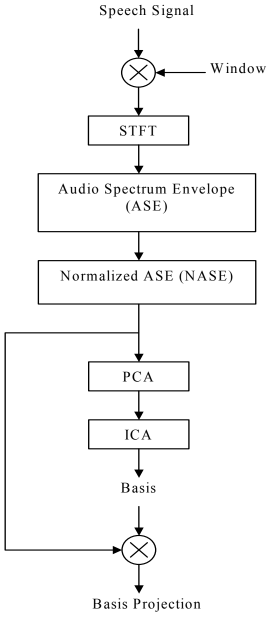

Figure 1: Block diagram of spectrum basis projection

We apply this general system to speech signals by calculating a basis for each speaker and then projecting the speaker's data onto his or her basis.

First, the observed speech signal $s(n)$ is divided into overlapping frames by the application of a hamming window function and analyzed using the short-time Fourier transform (STFT)

$$S(k,l)=\sum_{n=0}^{N-1} s(n+lM)w(n)\exp(-j(2\pi/N)nk), \quad (1)$$

where k is the frequency bin index, l is the time frame index, w is an analysis window of size lw, and M is the hop size. By Parseval's theorem (i.e., so that power is preserved) there is a further factor of $1/N$ to equate the sum of the squared magnitude of the STFT coefficients to the sum of the squared, zero-padded, windowed signal as

$$P(k,l)=\frac{1}{lw\cdot N}\left|S(k,l)\right|^2 \quad (2)$$

To extract reduced-rank spectral features, the spectral coefficients $P(k,l)$ are grouped in logarithmic sub-bands. Frequency channels are logarithmically spaced in non-overlapping ¼-octave bands spanning between 62.5 Hz, which is the default "low edge" and 8 kHz, which is the default "high edge". The output of the logarithmic frequency range is the weighted sum of the power spectrum in each logarithmic sub-band. The spectrum according to a logarithmic frequency scale, which the MPEG-7 standard refers to as Audio Spectrum Envelope (ASE), consists of one coefficient representing power between 0 Hz and low edge, a series of coefficients representing power in logarithmically spaced bands between low edge and high edge, and a coefficient representing power above high edge.

The resulting log-frequency power spectrum is converted to the decibel scale

$$D(f,l)=10\log_{10}(ASE(f,l)), \quad (3)$$

where f is the logarithmic frequency range.

Finally, each decibel-scale spectral vector is normalized with the RMS (root mean square) energy envelope, thus yielding a normalized log-power version of the ASE (NASE). The full-rank features for each frame l consist of both the RMS-norm gain value R_l and the normalized ASE (NASE) vector X_l:

$$R_l=\sqrt{\sum_{f=1}^{F}(10\log_{10}\{ASE(f,l)\})^2}, 1\le f\le F \quad (4)$$

and

$$X(f,l)=\frac{10\log_{10}\{ASE(f,l)\}}{R_l}, \quad 1\le l\le L \quad (5)$$

where F is the number of ASE spectral coefficients and L is the total number of frames.

To help the reader visualize the kind of information that the NASE vectors X_l convey, 3D-plots of the NASE of a male and a female speaker reading the sentence "Handwerker trugen ihn" are shown in Figure 2. In order to make the images look smoother, the frequency channels are spaced with 1/16-octave bands instead of the usual ¼-octave bands. The reader should note that recognizing the gender of the speaker by visual inspection of the plots is easy – compared to the female speaker, the male speaker produces more energy at the lower frequencies and less at the higher frequencies.

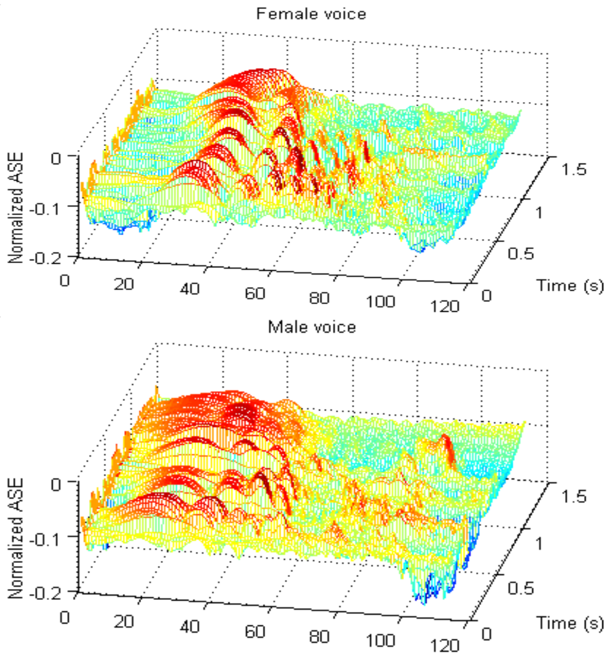

Figure 2: 3D-plots of the normalized ASE of a male speaker and a female speaker

The next step in the feature extraction is to extract a subspace using PCA from the NASE. Then, to yield a statistically independent component basis, the FastICA [7] algorithm is used. Some preprocessing is useful before using FastICA to estimate the un-mixing or uncorrelated matrix W. In the following, the rows represent the spectral vectors and the columns represent the time frames. First the rows should be centered by subtracting the mean value of each column from the column:

$$\hat{X}(f,l)=X(f,l)-\mu_f, \quad (6)$$

$$\mu_f=\frac{1}{L}\sum_{l=1}^{L}X(f,l),$$

(7)

where μ_f is the mean of the column *f*.

Standardizing the columns by making sure that the rows have no DC-offset and a unit variance is also a good idea:

$$\mu_l=\frac{1}{F}\sum_{f=1}^{F}\hat{X}(f,l) \quad (8)$$

$$\chi_l=\sum_{f=1}^{F}\hat{X}^2(f,l), \quad (9)$$

$$\Gamma_l=\sqrt{(\chi_l-F\cdot\mu_l^2)/(F-1)}, \quad (10)$$

$$\hat{X}(f,l) = \frac{\hat{X}(f,l) - \mu_l}{\Gamma_l}, \qquad (11)$$

where μ_l is the mean, χ_l is the energy of the whitened NASE and Γ_l is the standard deviation of the row *l*. In a further step the columns are whitened, which means that they are linearly transformed so that the components are uncorrelated and have unit variance. Whitening can be performed via eigenvalue decomposition of the covariance matrix

$$C = VDV^T = \mathrm{E}\left\{\hat{X}\,\hat{X}^T\right\}, \qquad (12)$$

$$C_P = D^{-\frac{1}{2}}V^T \qquad (13)$$

where V is the matrix of orthogonal eigenvectors and D is a diagonal matrix with the corresponding eigenvalues. In order to perform dimensionality reduction, we reduce the size of the matrix C_P by throwing away $F-E$ of the columns of C_P corresponding to the smallest eigenvalues of D. We call the resulting matrix C_E, which has the dimensions $F \times E$. The whitening is done by multiplication with the $F \times E$ transformation matrix C_E and $L \times F$ matrix $\hat{X}$:

$$\check{X} = \hat{X}\,C_E \qquad (14)$$

This method of whitening is closely related to PCA. After extracting the reduced PCA basis C_E, a further step consisting of basis rotation in the directions of maximal statistical independence is required for applications that require maximum separation of features, such as the separation of source components of a spectrogram. A statistically independent basis is derived using an additional step of ICA after PCA extraction. The input basis vectors are then fed to the FastICA algorithm, which maximizes the information with the following six steps:

1. Initialize spectrum basis W_i to small random values
2. Newton method

$$W_i = \mathrm{E}\left\{\check{X}\, g\left(W_i^T\,\check{X}\right)\right\} - \mathrm{E}\left\{g'\left(W_i^T\,\check{X}\right)\right\}W_i \qquad (15)$$

where g is the derivative of non-quadratic function.

3. Normalization

$$W_i = \frac{W_i}{\|W_i\|} \qquad (16)$$

4. De-correlation by Gram-Schmidt orthogonalization

$$W_i = W_i - \sum_{j=1}^{i-1} W_i^T W_j W_j \qquad (17)$$

5. Normalization

$$W_i = \frac{W_i}{\|W_i\|} \qquad (18)$$

6. If not converged, go back to step 2.

The purpose of the Gram-Schmidt decorrelation/orthogonalization performed in the algorithm is to avoid finding the same component more than once. When the tolerance becomes close to zero, the Newton method will usually keep converging towards that solution, and so by turning off the decorrelation when almost converged, the orthogonality constraint is loosened. Steps 1-6 are executed until convergence. Then the iteration performing only the Newton step and normalization are done until convergence $W_i W_i^T = 1$. With this modification the true maximum is found. The resulting spectrum projection is the product of the NASE matrix X, the dimension-reduced PCA basis functions C_E and the ICA transformation matrix W :

$$Y = XC_E W \qquad (19)$$

This spectrum projection is the compliant with the spectrum projection from the MPEG-7 standard and is used to represent low-dimensional features of a spectrum after projection onto a reduced rank basis.

3. Training using real data and recognition experiments

For speaker recognition, 25 speakers were used, 11 male and 14 female. Each speaker was instructed to read 15 different sentences. After we used a sampling rate of 22.05kHz to record the speakers reading the sentences, we cut the recordings into smaller clips: 16 training clips (about 60 seconds total), 5 additional longer training clips (60 s.), and 5 test clips (20 s.) per speaker. In order to determine if the amount of training data plays an important role for the different feature extraction methods, we defined two different training sets: the smaller set included only the 16 training clips and was 60 seconds long, and the larger set included the original 16 plus the 5 additional longer clips and was about 120 seconds long. Each speaker was modeled by a left-right HMM with 5 states. For each feature space (NASE, PCA, ICA), a set of 25 HMMs was trained using a classical Expectation and Maximization (EM) algorithm.

In the case of NASE, the matching process was easy because there were no bases. We simply matched each test clip against each of the 25 HMMs (trained with NASE features) via the Viterbi algorithm. The HMM yielding the best acoustic score (along the most probable state path) determined the recognized speaker.

In the case of the ICA and PCA methods, each HMM had been trained with data projected onto a basis as depicted in Figure 3. So, every time we tested a sound clip on an HMM, we had to first project the sound clip's NASE onto the HMM's basis. On the one hand, this process caused testing to last considerably longer, as each test clip had to be projected onto 25 different bases, before it could be tested on the 25 HMM's to determine what it should be recognized as, but on the other hand, the performance due to the projection onto the well-chosen bases increased performance considerably. In order to obtain good results with the PCA and ICA algorithms, feature extraction parameters needed to be selected with care.

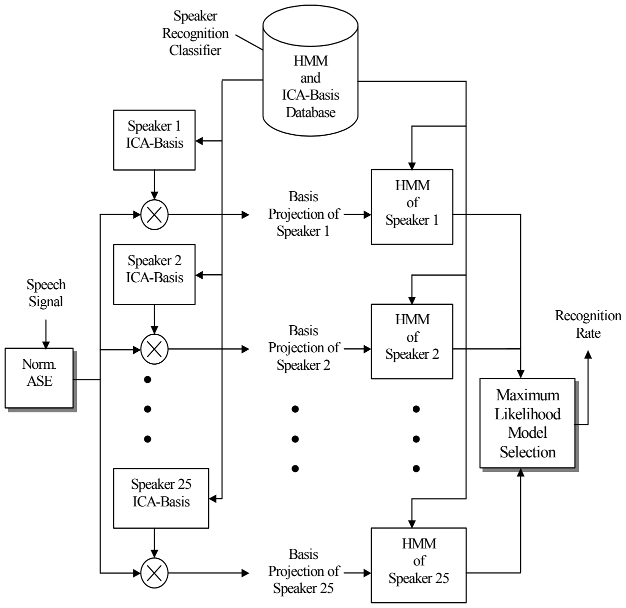
Speaker Recognition Classifier
HMM and ICA-Basis Database
Speaker 1 ICA-Basis
Speaker 2 ICA-Basis
Speaker 25 ICA-Basis
Speech Signal
Norm. ASE
Basis Projection of Speaker 1
Basis Projection of Speaker 2
Basis Projection of Speaker 25
HMM of Speaker 1
HMM of Speaker 2
HMM of Speaker 25
Recognition Rate
Maximum Likelihood Model Selection

Figure 3: Block diagram of the classification using spectrum basis projection features.

The parameter with the most drastic impact turned out to be the horizontal dimension E of the matrix C_E from PCA. If E was too small, the matrix C_E reduced the data too much, and the hidden Markov models did not receive enough information. However, if E became too large, then the extra information extracted was not very important and would have better been ignored. The recognition rate versus E from the PCA and ICA methods for the smaller training set are depicted in Figure 4:

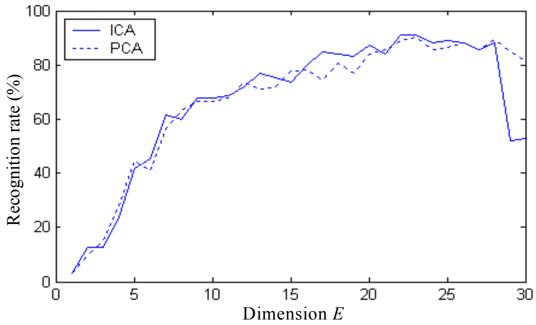

Figure 4: Comparison of recognition rate of PCA and ICA.

As can be seen above, the best value for both methods E was 23. However, this was not always the case. We also generated the plot for speaker recognition among 6 male speakers, which revealed that the optimal dimension for E should be 16, so it seems that one needs to be careful about choosing E and might have to test empirically to find the optimal value.

The results of our tests using the different feature extraction methods are shown in Table 1. For PCA and ICA we simply took the recognition rate corresponding to $E = 23$, even though in one case the recognition rate was 1.5% higher for $E = 28$ (PCA, with larger training set).

Regarding the recognition of 25 speakers, ICA yields better performance than PCA and NASE features. The resulting recognition rates (90.4-93.6%) using MPEG-7 conform audio descriptors appear to be slightly lower than the 98% recognition rate that we obtained with classical Mel-Frequency Cepstral Coefficients (MFCC). However, many applications that are likely to employ MPEG-7 descriptions in the future are not security applications and may thus not necessarily require very high speaker recognition rates.

To test gender recognition, we used the smaller set. Two HMMs were trained: one with the training clips from female speakers, the other with the training clips from male speakers. Because there were only two possible answers to the recognition question: male or female, this experiment was naturally much easier to carry out and resulted in excellent recognition rates, as depicted in Table 1. 100% indicates that 0 mistakes were made out of 125 test sound clips.

Table1: Comparison of speaker recognition accuracies (%) between several feature extraction methods.

Recognition Mode	Norm. ASE	PCA	ICA
Speaker recognition (small set)	80.8	90.4	91.2
Speaker recognition (larger set)	80	85.6	93.6
Gender recognition (small set)	98.4	100	100

4. Conclusions

In this paper, the use of spectrum basis projection features based on the MPEG-7 standard for the purpose of speaker recognition were introduced and analyzed. The spectrum basis functions were computed from the NASE using a basis decomposition algorithm such as PCA and FastICA. An experimental 25 speaker recognition system was implemented using reduced-rank projection features (in conformance with the MPEG-7 standard) and HMM classifiers. The ICA basis projection features demonstrated better speaker and gender recognition performance than the NASE features and the PCA basis projection features.

5. References

[1] Jolliffe, I. T., "Principal component analysis", *Springer-Verlag*, 1986.

[2] Comon, P., "Independent component analysis, A new concept?", *Neural Computation*. Vol. 7, no.6, pp. 1004-1034, 1995.

[3] Huang, C., Chen, T., Li, S., Chang, E., Zhou, J., "Analysis of speaker variability", *Eurospeech*, pp. 1377-1380, Sep. 2001.

[4] Manjunath, B., Salembier, P., Sikora, T., "Introduction to MPEG-7", *Willey*, April 2002.

[5] Casey, M.A., "General Sound Similarity and Sound Recognition Tools", in "Introduction to MPEG-7", *Willey*, April 2002.

[6] ISO/IEC JTC 1/SC 29, "Information technology multimedia content description interface-Part 4: Audio", June, 2001.

[7] Hyvarinen, A., Oja, E., "Independent component analysis: algorithms and application", *Neural Networks13*. pp. 411-430, 2000.