Comparison of MPEG-7 Basis Projection Features and MFCC applied to Robust Speaker Recognition

Hyoung-Gook Kim, Martin Haller, Thomas Sikora

Department of Communication Systems
Technical University of Berlin, Germany
{kim,haller,sikora}@nue.tu-berlin.de

Abstract

Our purpose is to evaluate the efficiency of MPEG-7 basis projection (BP) features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC) for speaker recognition in noisy environments. The MPEG-7 feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE), a basis decomposition algorithm and a spectrum basis projection. Prior to the feature extraction the noise reduction algorithm is performed by using a modified log spectral amplitude speech estimator (LSA) and a minima controlled noise estimation (MC). The noise-reduced features can be effectively used in a HMM-based recognition system. The performance is measured by the segmental signal-to-noise ratio, and the recognition results of the MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC) in comparison to other noise reduction methods. Results show that the MFCC features yield better performance compared to MPEG-7 features.

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech signals. 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. The dominant features used for speech/speaker recognition are Mel-scale Frequency Cepstrum Coefficients (MFCC).

Recently the MPEG-7 international standard has adopted dimension-reduced, decorrelated spectral features based on independent component analysis (ICA) [1] basis functions for general sound recognition framework [2] using hidden Markov models (HMM) [3]. Many researchers are interested in comparing of the performance of MPEG-7 ASP features vs. MFCCs according to reduced dimension.

In [4], we implemented and analyzed the MPEG-7 basis projection (BP) features for the purpose of a speaker recognition system. We compared the results of the HMM classification with each of the extracted feature sets.

In many speech communication applications, like audio-conferencing, hands-free mobile telephony and speech/speaker recognition devices, the recorded speech signals contain a considerable amount of acoustic noise, which not only degrades the subjective speech quality, but also hinders performance of recognition systems. Therefore efficient noise reduction algorithms are called for.

In this paper, the noise-reduced basis projection features are applied to speaker recognition system in noisy environments. The noise reduction algorithm is based on a simple modified log spectral amplitude speech estimator (MLSA) and a minima controlled (MC) noise estimation. For the measure of the performance we compare the recognition results of the MPEG-7 basis projection features vs. MFCC.

2. Noise reduction preprocessing

The noise reduction preprocessing in combination with feature extraction improve both speech recognition performance and the quality of speech reconstruction under noisy conditions.

The noise reduction algorithm based on simple modified LSA speech estimator and a minima controlled noise estimation (MLSA-MC) shown in Figure 1 operates in the frequency domain. A nonlinear frequency dependent gain function is applied to the spectral components of the noisy speech signal in an attempt to obtain estimates of spectral components of corresponding clean speech.

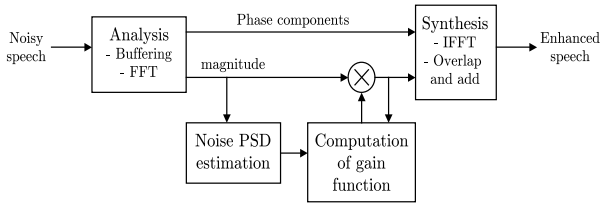

Figure 1: *Block diagram of the noise reduction.*

Let $S(k) = A(k)\exp(j\varphi(k))$, $D(k)$, $Y(k) = R(k)\exp(j\theta(k))$ be the Fourier expansions of clean speech $s(n)$, additive noise $d(n)$, and noisy speech respectively $y(n)$.

The estimation of clean speech is obtained by applying a modified log-spectral amplitude gain function G_L to each spectral component of the noisy speech signal:

$$\begin{aligned} O(k,l) &= [G_{LSA}(k,l)]^{G_M(k,l)}\, R(k,l) \\ &= G_L(k,l)R(k,l)\,, \end{aligned} \tag{1}$$

where G_M is the gain modification function and G_{LSA} is the gain function. G_M is applied to take into account the probability of the speech presence in the frequency k, and it is referred to as soft-decision modification of the optimal estimation. G_M is given by

$$G_M(k,l) = \frac{\Lambda(k,l)}{1+\Lambda(k,l)}\,, \tag{2}$$

where $\Lambda(k,l)$ is a likelihood ratio between speech presence and

speech absence in frequency k and defined by

$$\Lambda(k,l) = \frac{1-q(k,l)}{q(k,l)} \left. \frac{\exp\left(v(k,l)\right)}{1+\xi(k,l)} \right|_{\xi(k,l)=\frac{\eta(k,l)}{1-q(k,l)}} . \quad (3)$$

The G_{LSA} is derived by

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(0.5 \int_{t=v(k,l)}^{\infty} \frac{\mathrm{e}^{-t}}{t} dt \right), \quad (4)$$

where the a posteriori SNR $\gamma(k,l)$, $\xi(k,l)$ and $v(k,l)$ are defined as

$$\gamma(k,l) = \frac{R^2(k,l)}{\lambda_d(k,l)} , \quad (5)$$

$$\xi(k,l) = \frac{\eta(k,l)}{1-q(k,l)} \text{ , and} \quad (6)$$

$$v(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \gamma(k,l) . \quad (7)$$

The a priori SNR $\eta(k,l)$ is defined as

$$\begin{aligned} \eta(k,l) = \beta G_L^2(k,l-1)\gamma(k,l-1) + \\ (1-\beta)\max\left\{\gamma(k,l)-1\right\} \end{aligned} \quad (8)$$

where β ($0 < \beta < 1$) is the SNR smoothing factor, $q(k,l)$ is an estimate of speech absence a priori probability and $\lambda_d(k,l)$ is a noise spectrum estimate. The amount of noise reduction can be reduced by overestimation $\eta(k,l)$ and increased by underestimating $\eta(k,l)$.

The a priori speech absence probability $q(k,l)$ is estimated by

$$q(k,l) = bq(k,l-1) + (1-b)U(k,l) , \quad (9)$$

where b ($0 < b < 1$) is time-smoothing factor and the likelihood ratio

$$U(k,l) = \begin{cases} 1 & \text{if } \gamma(k,l) > \zeta_1 \\ \frac{\zeta_1 - \gamma(k,l)}{\zeta_1 - \zeta_2} & \text{if } \zeta_2 < \gamma(k,l) < \zeta_1 \\ 0 & \text{otherwise} \end{cases} . \quad (10)$$

The noise estimate is obtained by a careful balance between spectral power values and its minimum tracking $M(k,l)$, using a time-varying frequency-dependent smoothing parameter that is adjusted by the speech presence probability in adverse environments involving non-stationary noise, weak speech components and low input SNR:

$$\lambda_d(k,l) = \alpha_d(k,l)M(k,l) + \left(1-\alpha_d(k,l)\right) R^2(k,l) \quad (11)$$

using a smoothing parameter

$$\alpha_d(k,l) = \alpha + (1-\alpha)p(k,l) \quad (12)$$

with time-varying smoothing parameter α ($0 < \alpha < 1$).

The conditional speech presence probability is defined as

$$p(k,l) = \alpha_p p(k,l-1) + (1-\alpha_p)I(k,l) \quad (13)$$

with time-varying smoothing parameter α_p ($0 < \alpha_p < 1$).

The minimum values $M(k,l)$ of an average $E(k,l)$ of the short-time noisy spectrum are calculated within windows of S frames whether speech is present or not. The minimum value for the current frame is found by a comparison with the stored minimum value:

$$M(k,l) = \min_{s=0\ldots S} \left\{ M(k,l-s), E(k,l) \right\} , \quad (14)$$

where the average of the short-time noisy spectrum

$$E(k,l) = \frac{1}{B} \sum_{l=0}^{B-1} P(k,l) . \quad (15)$$

$$P(k,l) = \sum_{m=-1}^{1} b(i) R^2(k-m,l) \quad (16)$$

is performed over B frames and a window function b in frequency.

The indicator function $I(k,l)$ for the voice activity detector is defined by

$$I(k,l) = \begin{cases} 1 & \text{if } \frac{E(k,l)}{M(k,l)} > \psi \\ 0 & \text{otherwise} \end{cases} . \quad (17)$$

3. Feature extraction using basis projection

The MPEG-7 feature extraction mainly consists of a Normalized Audio Spectrum Envelope (NASE), a basis decomposition algorithm and a spectrum basis projection.

3.1. Normalized Audio Spectrum Envelope (NASE)

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

$$S(k,l) = \sum_{n=0}^{N-1} s(n+lM)w(n) \exp\left(\frac{-j2\pi nk}{N} \right) \quad (18)$$

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 magnitudes of the STFT coefficients as

$$P(k,l) = \frac{1}{nfN} \left| S(k,l) \right|^2 \quad (19)$$

where the window normalization factor nf is defined as

$$nf = \sum_{n=0}^{lw-1} w^2(n) . \quad (20)$$

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 octave bands spanning between *low edge* and *high edge*. The output of the logarithmic frequency range is the weighted sum of the power spectrum in each logarithmic sub-band. The resulting log-frequency power spectrum is converted to the decibel scale

$$D(f,l) = 10\log_{10}\left(ASE(f,l) \right) , \quad (21)$$

where f ($f < k$) 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 i consist of both the RMS-norm gain value R_l and the normalized ASE (NASE) vector X_l:

$$R_l = \sqrt{\sum_{f=1}^{F} \left[10\log_{10}\left(ASE(f,l) \right) \right]^2} , 1 \leqslant f \leqslant F , \quad (22)$$

and

$$X(f,l) = \frac{10\log_{10}\left(ASE(f,l)\right)}{R_l}\,,\, 1 \leqslant l \leqslant L\,, \qquad (23)$$

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

3.2. Projection features onto basis decomposition algorithm

The next step in the feature extraction is to extract a subspace using PCA/ICA from the NASE matrix. Then, to yield a statistically independent component basis, the FastICA [1] algorithm is used.

Before FastICA algorithm, whitening closely related to PCA is performed via eigenvalue decomposition of the covariance matrix,

$$C = VDV^T = \mathrm{E}\left\{\hat{X}\hat{X}^T\right\}\,, \qquad (24)$$

$$C_P = D^{-\frac{1}{2}}V^T\,, \qquad (25)$$

where $\hat{X}$ is the centered data from X, V is a 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 X_E:

$$X_E = \hat{X}C_E \qquad (26)$$

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 requiring maximum separation of features. 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

$$W_i = \mathrm{E}\left\{\check{X}g\left(W_i^T\check{X}\right)\right\} - \mathrm{E}\left\{g'\left(W_i^T\check{X}\right)\right\}W_i \qquad (27)$$

   where g is the derivative of non-quadratic function.
3. Normalization $W_i = W_i/\|W_i\|$
4. De-correlation by Gram-Schmidt orthogonalization

$$W_i = W_i - \sum_{j=1}^{i-1} W_i^T W_j W_j \qquad (28)$$

5. Normalization $W_i = W_i/\|W_i\|$
6. If not converged, go back to step 2.

Steps 1-6 are executed until convergence. Then the iteration performing only the Newton step and normalization are done until convergence $W_i W_i^T = 1$. 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:

$$\tilde{X} = XC_EW \qquad (29)$$

4. Experiments

For the performance of the noise reduction, we measure segmental SNR improvement in speech segments and speech recognition rate. For the speaker recognition we compare the recognition results of MPEG-7 standardized features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC) concatenated with noise reduction algorithm.

4.1. Segmental SNR improvement

To measure the performance of the proposed noise reduction algorithm in comparison to other one-channel noise reduction methods, the segmental signal-to-noise ratio (seg.SNR) is computed by $improveSNR = seg.SNR_{out} - seg.SNR_{in}$ for the enhanced speech signals.

Three types of background noise - white noise, car noise and factory noise - were artificially added to different portions of the data at SNR of 10 dB and 5 dB. The speech data used for the segmental SNR improvement were digitized at 22.05 kHz using 16 bits per sample.

Table 1 shows that our MLSA-MC algorithm gives best improvement results compared to the results of MM-LSA (multiplicatively modified log-spectral amplitude speech estimator) [5] and OM-LSA (optimally modified LSA speech estimator and minima controlled recursive averaging noise estimation) [6].

Table 1: *Comparison of segmental SNR improvement of different one-channel noise estimation methods.*

methods	Input SNR [dB]					
	white noise		car noise		factory noise	
	10	5	10	5	10	5
MM-LSA	7.3	8.4	8.2	9.7	6.2	7.7
OM-LSA	7.9	9.9	9	10.6	6.9	8.3
MLSA-MC	8.1	10	9.2	10.8	7.1	8.4

4.2. Recognition accuracy on noisy TI digits database

For evaluation of the improvement of speech recognition with the noise reduction algorithms, the Aurora 2 database together with a HTK software tools has been chosen and the multi-condition training on noisy data is used.

The feature vector from the speech database with sampling rate 8 kHz consists of 39 parameters: 13 mel frequency cepstral coefficients plus delta and acceleration calculations. The mel-cepstrum coefficients were modeled by a simple left-to-right 16-state three-mixture whole word HMM. For the noisy speech results, we averaged the word accuracies between 0 dB and 20 dB SNR.

In the Table 2, set A, B, and C refer to matched noise condition, mismatched noise condition, and mismatched noise and channel condition, respectively. Table 2 describes the results of the recognition accuracy.

As seen in the results of Table 2, MLSA-MC provides slightly better performance than MM-LSA front-end and is similar to OM-LSA front-end. However, the MLSA-MC method is more simple and does not need more computational cost compared to OM-LSA.

Table 2: *Comparisons of word accuracies (%) between several front-ends on the Aurora 2 database.* NR: *noise reduction*

Feature Extraction	Set A	Set B	Set C	Overall
Without NR	87.81	86.27	83.77	85.95
MM-LSA	89.68	88.43	86.81	88.30
OM-LSA	90.93	89.48	88.92	89.77
MLSA-MC	91.35	89.65	89.09	90.03

4.3. Results of speaker recognition

For speaker recognition we performed experiments where 25 speakers (11 male and 14 female) were used. Each speaker was instructed to read 15 different sentences. After recording of the sentences spoken by each speaker, we cut the recordings into smaller clips: 21 training clips (about 3 minutes long), and 10 test clips (50 s) per speaker.

The speech data were digitized at 22.05 kHz using 16 bits per sample. The non-stationary Gaussian white noise is artificially added to the speech database at the SNR ratios ranging from clean over 20 dB to 5 dB in steps of 5 dB.

The ASP features based on PCA/ICA basis were derived from speech frames of length 30 ms with a frame rate of 15 ms. The lower and upper boundary of the logarithmic frequency bands are 62.5 Hz and 8 kHz that are over a spectrum of 7 octaves.

The training phase is done only on a clean speech signal without noise reduction preprocessing, while test phase is done on a noisy speech signal (mismatched conditions).

In order to compare the performance of MPEG-7 standardized features vs. MFCC approach for speaker recognition we used MFCC coefficients without delta and acceleration calculations. For the recognizer a 7-state left-right HMM model were applied.

The results of speaker recognition are shown in Table 3. For decreasing SNR ratios, the speaker recognition rate without noise reduction is seriously decreased due to mismatched conditions. In noisy environments the MPEG-7 ASP, which is projected onto PCA/ICA basis from the clean speech training database, yields better performance than MFCC.

However, the concatenation of MFCC with noise reduction using MLSA-MC yields a higher recognition rate than the concatenation of MPEG-7 ASP features with MLSA-MC. Overall the features with noise reduction based on MLSA-MC significantly improve the speaker recognition system performance.

5. Conclusions

In this paper, we evaluate the efficiency of MPEG-7 dimension-reduced, decorrelated log-spectral features vs. Mel-scale Frequency Cepstrum Coefficients (MFCC) for the speaker recognition in noisy environments. Our results show that the MFCC features yield better performance compared to MPEG-7 ASP in combination with the noise reduction pre-processing on noisy speaker recognition task. In the case of MFCC, the process of feature extraction is simple and fast while the extraction of the MPEG-7 ASP is more time and memory consuming compared to MFCC.

For our further research, we will compare the performance of MPEG-7 ASP features based on several basis decomposition algorithms vs. cepstrum-domain feature compensation.

Table 3: *Comparison of speaker recognition accuracies (%) between several feature extraction methods.* FE: *feature extraction*, NR: *noise reduction*, Cl: *clean speech*, P: *basis projection based on* PCA, I: *basis projection based on* ICA, M: MFCC, N: *noise reduction.*

(a) Recognition accuracy with feature dimension 7

FE	Speech Material				
	Cl.	Noisy speech			
		20 dB	15 dB	10 dB	5 dB
P	65.8	24.9	17.3	10.2	7.6
P+N	50.7	55.1	51.6	36	22.2
I	66.2	27.6	16.4	11.1	7.6
I+N	55.6	58.2	53.8	38.2	21.3
M	71.1	23.6	17.8	10.7	7.6
M+N	63.6	58.2	51.1	37.3	21.8

(b) Recognition accuracy with feature dimension 13

FE	Speech Material				
	Cl.	Noisy speech			
		20 dB	15 dB	10 dB	5 dB
P	83.6	48.9	42.7	32.9	27.6
P+N	82.2	76	65.8	49.8	38.7
I	84.9	47.6	41.3	34.2	24.9
I+N	82.2	77.8	67.1	56.9	37.3
M	91.6	50.7	24.9	11.6	4.4
M+N	90.7	81.8	78.2	64.4	39.6

(c) Recognition accuracy with feature dimension 23

FE	Speech Material				
	Cl.	Noisy speech			
		20 dB	15 dB	10 dB	5 dB
P	90.2	61.3	48	36.9	28.9
P+N	92.9	80.9	75.1	56.4	43.1
I	92.9	66.2	53.3	39.6	28.9
I+N	94.7	82.2	78.2	64	49.3
M	95.6	66.2	37.8	24	19.6
M+N	94.7	92.4	89.3	82.2	57.8

6. References

[1] Hyvarinen, A., and Oja, E., "Independent component analysis: algorithms and applications", Neural Networks, Vol. 13, 2000, p 411-430.

[2] Casey, M., "MPEG-7 sound recognition tools", IEEE Trans. Circuits and Systems for Video Technology, Vol. 11, 2001.

[3] Rabiner, L., and Juang, B.-H., "Fundamentals of speech recognition", Prentice Hall, N.J., 1993.

[4] Kim, H.-G., Berdahl, E., Moreau, N., and Sikora, T., "Speaker recognition using MPEG-7 descriptors", Eurospeech 2003, September 2003.

[5] Malah, D., Cox, R.V., and Accardi, A.J., "Tracking speech presence uncertainty to improve speech enahnacement in non-stationary noise environments", ICASSP 1999, p 789-792.

[6] Cohen, I., and Berdugo, B., "Speech enhancement for non-stationary noise environments", Signal Processing, Elsevier, Vol. 81, 2001, p 2403-2418.