Simple Noise Robust Feature Vector Selection Method for Speaker Recognition

Gabriel Hernández, José R. Calvo, Flavio J. Reyes, and Rafael Fernández

Advanced Technologies Application Center
{gsierra,jcalvo,freyes,rfernandez}@cenatav.co.cu
http://www.cenatav.co.cu

Abstract. The effect of additive noise in a speaker recognition system is known to be a crucial problem in real life applications. In a speaker recognition system, if the test utterance is corrupted by any type of noise, the performance of the system notoriously degrades. The use of a feature vector selection to determine which speech frames are less affected by noise is the purpose in this work. The selection is implemented using the euclidean distance between the Mel features vectors. Results reflect better performance of robust speaker recognition based on selected feature vector, as opposed to unselected ones, in front of additive noise.

Keywords: speaker verification, cepstral features, selected feature vector, channel mismatch.

1 Introduction

Speech signal varies due to differences introduced by microphone, telephone, gender, age, and other factors, but a key problem is the presence of noise in the signal, which can provoke an awful performance in the speech processing algorithms working under extreme noisy conditions. Wireless communications, digital hearing aids or robust speech recognition, are examples of such systems which frequently require a noise reduction technique.

Recently, much research has been conducted in order to reduce the effect of handset/channel mismatch in speech and speaker recognition. Linear and nonlinear compensation techniques have been proposed, in the (a) feature, (b) model and (c) match-score domains [1]:

(a) Feature compensation methods [2]: filtering techniques such as cepstral mean subtraction or RASTA, discriminative feature design, and other feature transformation methods such as affine transformation, magnitude normalization, feature warping and short time Gaussianization.

(b) Model compensation methods [3]: speaker-independent variance transformation, speaker models transformation from multi-channel training data, and model adaptation methods.

(c) Score compensation methods [4]: aims to remove handset-dependent biases from the likelihood ratio scores as, H-norm, Z-norm, and T-norm.

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 313–320, 2009.

Other methods to reduce specifically the impact of noise have been proposed [1]:

- filtering techniques,
- noise compensation,
- use of microphone arrays and,
- missing-feature approaches.

The features most commonly used are: static and dynamic Mel Frequency Cepstral Coefficients (MFCC), energy, zero crossing rate and pitch frequency. The classification methods commonly used are: Frame and utterance energy threshold, noise level tracking or model based. This paper investigates a feature vector selection method over MFCC in speaker recognition, using speech samples distorted by noise. This features vectors are selected by mean of clustering of the MFCC using as criterion of Euclidean distance. To evaluate the selection method the Gaussian Mixture Model (GMM) [5] is used as baseline.

The rest of the paper is organized as follows. Section 2 explains the sensitivity of the Gaussian components. Section 3 describes the feature vector selection algorithm. Section 4 shows the results of the experiments. Finally section 5 presents the conclusions and future work.

1.1 Sensitivity of the Gaussian Components

The GMM models the feature vectors of a speech signal, performing a weighted sum of M (number of mixtures of Gaussian probability density functions).

$$p(\hat{x}/\lambda) = \sum_{i=1}^{M} p_i b_i(\hat{x}), \tag{1}$$

We take as input data the MFCC.

$$MFCCMatrix \rightarrow X = \left\{ \begin{array}{cccc} \hat{x}_1 & \hat{x}_2 & \dots & \hat{x}_T \\ \downarrow & \downarrow & & \downarrow \\ c_{1,1} & c_{1,2} & \dots & c_{1,T} \\ c_{2,1} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ c_{D,1} & \dots & \dots & c_{D,T} \end{array} \right\} \tag{2}$$

where $\hat{x}_t$ is a feature vector that represents one observation over the signal, t is the index of the speech frame and D is the amount of coefficients. The matrix X is a sequence of random variables indexed by a discrete variable, time ($t = 1, \cdots, T$). Each of the random variables of the process has its own probability distribution function and we assume that they are independent. This is called MFCC matrix and is extracted from a speech expression, which characterizes the speaker.

where: $b_i(\hat{x}_t) \rightarrow$ with $i = 1, \cdots, M$ are the Gaussian density components. Each component is a Gaussian function of the form:

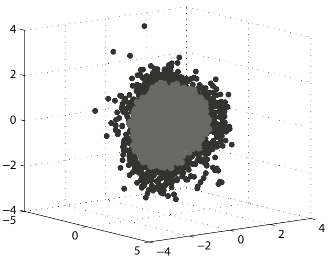

Fig. 1. First three MFCC

$$b_i(\hat{x}) = \frac{1}{(2\Pi)^{\frac{D}{2}}|\Sigma_i^{\frac{1}{2}}|} exp\{-\frac{1}{2}(\hat{x} - \hat{\mu}_i)' \Sigma_i^{-1}(\hat{x} - \hat{\mu}_i)\} \tag{3}$$

where: μ → mean matrix, Σ → covariance matrix.

and: p_i → weights of the mixtures $i = 1, 2, \cdots, M$ and satisfies that $\sum_{i=1}^{M} = 1$.

If we represent the first three MFCC (c1, c2, c3) to view their behavior, we would observe a very dense cloud of points toward the center and with some scattered at the edges, the same behavior will follow any group of coefficients that are chosen, for example:

If we generalize the representation of Fig. 1 to D MFCC, we could assume that the D representation would have a very dense cloud of points toward its $D - dimensional$ center.

What would happen if we classify a two dimensional data (c1, c2) or three dimensional data (c1, c2, c3) using 16 Gaussian mixtures?

For the graphical representation of 16 GMM of three MFCC we can conclude:

From this intuitive idea that the individual components of a multi-modal density (GMM) is capable of modeling the underlying acoustic classes in the speech for each speaker, and that speaker's acoustic space can be approximated by a set of acoustic classes (mixtures), we can observe that acoustic classes are more overlapping in the region where the features are more compact. This overlapping reduces the discriminative power of these acoustic classes.

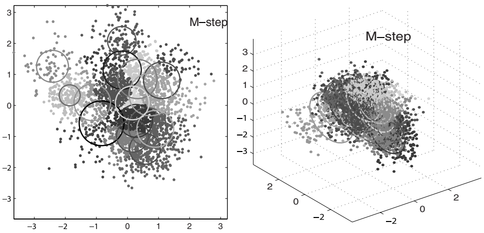

Fig. 2. Classification using 16 Gaussian mixtures of two and three dimensional data

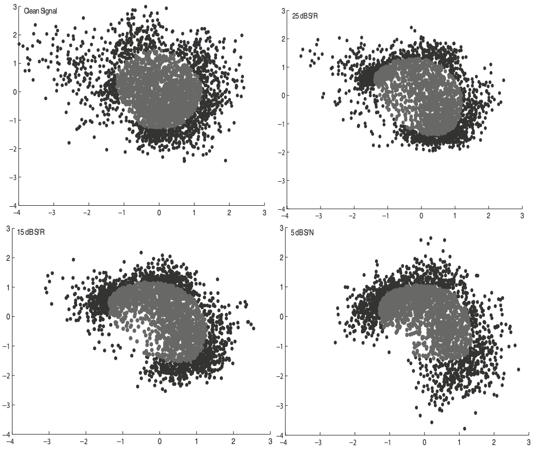

Fig. 3. Representation of the MFCCs distorted by additive white noise: Clean signal, 25db S/N, 15db S/N and 5db S/N

Gaussian components that define the cluster in the dense center of the features are much more overlapped between them, that the Gaussian components that define the features in the border. This makes the probability of Gaussian components given by the features in the dense center more prone to perturbations by the displacements of the features, in presence of noise, as is observed in Fig. 3.

The Fig. 3 shows what happens with the two dimensional coefficients when it is distorted by additive white noise, where it can be clearly seen how the points at the center of the cloud are affected on a larger scale. The intuitive idea that the individual components with less overlapping are capable of modeling the acoustic classes with more robustness in front of the displacements of the features in a noisy speaker verification process motivated us to find an algorithm of feature vector selection capable of only choosing those feature vectors that do not belong to the dense center.

2 Feature Selection Algorithms

From the above we developed an algorithm to select the feature vectors that are outside the dense center to use only these features in the verification process.

We use as input features the MFCC matrix $X_{D,T}$, assuming it describes the speech, then we take each as a point of acoustic space as shown in Fig. 1.

The algorithm can be summarized in four steps:

1. Construct a neighborhood graph - Compute its k neighbours more distant on the acoustic space based on Euclidean distances $d(i,j)$ between pairs of points i, j.

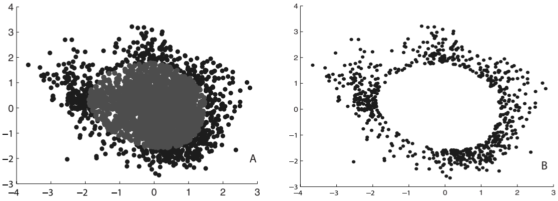

Fig. 4. Results of the feature vector selection (B)

2. Assign weights $W_{i,j}$ to the edges of the graph, in our case, in the $i-th$ row, will have the value 1 those points that belong to the neighboring of the point i, and will have the value 0 those points that are outside these neighboring.
3. Building an array L with length equal to amount of points (T) and in each index is stored the sum of the connections for this node, $y_j = \sum_i W_{i,j}$.
4. Select from the MFCC matrix only the features vectors that correspond to the indices of the previous array (L) which are different from zero. $\overline{X}_{D,V} = X_{D,T}$, where $V << T$.

The Fig. 4A shows the 3000 features vectors of the MFCC matrix. The clusters were defined with 16 neighbors. The Fig. 4B shows selected 826 << 3000 features vectors, all located at the edges, the features vectors located at the dense center were eliminated.

3 Experiments and Results

Ahumada [6] is a speech database of 103 Spanish male speakers, designed and acquired under controlled conditions for speaker characterization and identification. Each speaker in the database expresses six types of utterances in seven microphone sessions and three telephone sessions, with a time interval between them.

The experiment consisted in the evaluation of the performance of feature vector selection in speaker recognition, in front of noisy environment and channel mismatch using spontaneous phrases of 100 speakers in two telephones sessions of Ahumada. The white noise, hf-channel noise and pink noise obtained from Noisex database are artificially added to the samples at SNR ranking from 25 dB, 15 dB and 5 dB.

In order to evaluate the effectiveness of the proposed feature vector selection in speaker recognition, two recognition experiments were implemented for each type of noise and each SNR ranking using 12-dimensional MFCC + delta features vector with Cepstral Mean and Variance normalization applied.

1. Baseline experiment: train with 1 min of spontaneous sentences and test with the 1 min segments, with the SNR ranking applied.
2. Feature vector Selection experiment: the same baseline experiment but in the test phase feature vectors were selected using the proposed method.

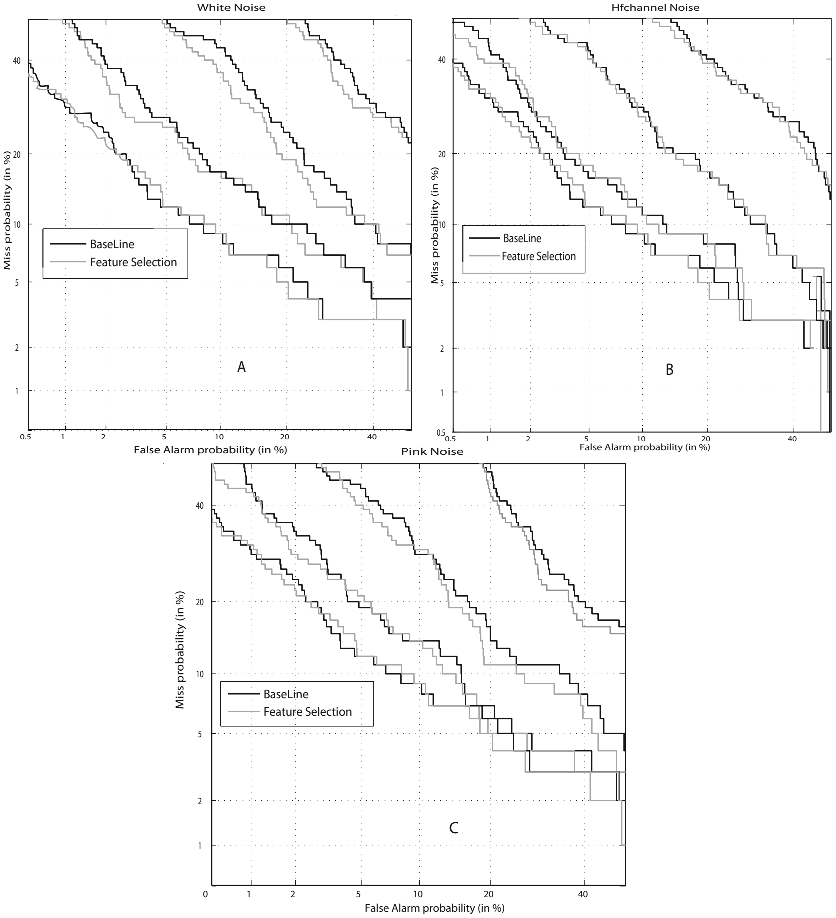

Fig. 5. Speaker recognition DET plots. Black: Baseline. Gray: Feature vector Selection. Clean Signal, 25 dB, 15 dB and 5 dB in the same order that increase the ERR. A) white noise, B) hf-channel noise and C) pink noise.

The performance of both experiments was evaluated using a 16 mixtures GMM classifier [5]. The results of the two experiments are reflected in detection error tradeoff (DET) plot [7], in Fig. 5:

The EER result of the experiments is shown in Table 1.

In the case of white noise (Fig. 5-A), for the first and second DET plot the result are alike, but in the third and fourth DET plot with 15 dB and 5 dB of S/N respectively is appreciable a better behavior using the feature vector selection. In the case of hf-channel noise (Fig. 5-B) the curves are similar in the two experiments for all levels of noise. In the case of pink noise (Fig. 5-C) the results are analogous to white noise, though in the second DET plot we start to view an improvement in the result using the feature vector selection, is valid to note that the white and pink noises are relatives.

Table 1. EER of the experiments

White Noise				
	Signal clean	25 dB SNR	15 dB SNR	5 dB SNR
Baseline	6	15	23	38
Feature Selection	8	15	19	32
Hf-channel Noise				
Baseline	6	13	18	34
Feature Selection	8	10	16	34
Pink Noise				
Baseline	6	15	19	33
Feature Selection	8	13	17	30

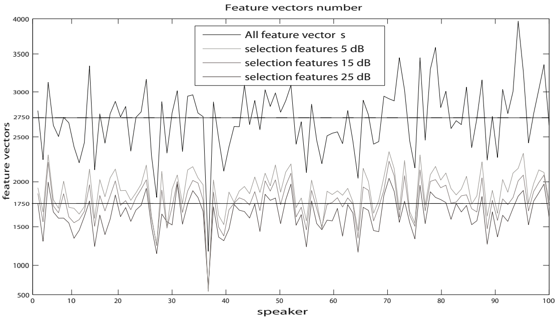

Fig. 6. Number of feature vectors for each speaker and their reduction after making the selection

It empirically shows that the features selected outside dense center are more noise robust for speaker recognition than all features vectors of the MFCC matrix, furthermore this selection allows a smaller amount of feature vectors for recognition. The Fig. 6 shows the difference in the amount of features.

Approximately, 1000 features vectors are eliminated by the selection in each speaker, which represents 20 seconds of each signal; this selection reduces the time calculation of the verification algorithms.

4 Conclusions and Future Work

The experiments results reflect a superior performance of selected MFCC respect to use all the MFCC in speaker recognition using speech samples from telephone sessions of Ahumada Spanish database.

- Results show that speaker recognition in noiseless conditions has the same behavior using either all MFCC or selected MFCC, but with increased noise selected feature show more robustness.
- Tests under noisy conditions (experiments A and C, 15 dB and 25 dB) reflect a better behavior of the selected feature respect to use all MFCC in front of

worst mismatch conditions, (channel and session variability) whereas in the experiment B have a similar behavior.

- In all experiments using the selected feature vectors the computation time of verification algorithms is reduced because of the elimination of 20 secs. from the complete signal.
- Experiments (Table 1) show an EER reduction due to utilization of selected feature vectors instead all MFCC. This reduction is 6 percent in high noisy conditions.

Future work will be in the direction of evaluate the influence of selected feature vectors in other noisy environments.

References

1. Ming, J., Hazen Timothy, J., Glass James, R., Reynolds Douglas, A.: Robust Speaker Recognition in Noisy Conditions. IEEE Trans. on ASLP 15(5) (July 2007)
2. Reynolds, D.A.: Channel robust speaker verication via feature mapping. Proc. of ICASSP, pp. II-53-6 (2003)
3. Teunen, R., Shahshahani, B., Heck, L.: A model-based transformational approach to robust speaker recognition. In: Proc. of ICSLP (2000)
4. Fauve, B.G.B., Matrouf, D., Scheffer, N., Bonastre, J.-F., Mason, J.S.D.: State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software. IEEE Trans. on ASLP 15(7), 1960–1968 (2007)
5. Douglas, A., Richard, R.y., Rose, C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. on SAP 3(1) (January 1995)
6. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V.: AHUMADA A large speech corpus in Spanish for speaker characterization and identification. Speech communication (31), 255–264 (2000)
7. Martin, A., et al.: The DET curve assessment of detection task performance. Proc. of EuroSpeech 4, 1895–1898 (1997)