Speaker Verification in Noisy Environment Using Missing Feature Approach

Dayana Ribas[1], Jesús A. Villalba[2], Eduardo Lleida[2], and José R. Calvo[1]

[1] Advanced Technologies Application Center (CENATAV), 7a ♯ 21812 e/ 218 y 222,
Rpto. Siboney, Playa, C.P. 12200, La Habana, Cuba
[2] Communications Technology Group (GTC), Aragon Institute for Engineering
Research (I3A), University of Zaragoza, Spain
{dribas,jcalvo}@cenatav.co.cu, {villalba,lleida}@cenatav.co.cu

Abstract. In order to handle speech signals corrupted by noise in speaker verification and provide robustness to systems, this paper evaluates the use of missing feature (MF) approach with a novel combination of techniques. A mask estimation based on spectral subtraction is used to determine the reliability of spectral components in a speech signal corrupted by noise. A cluster based reconstruction technique is used to remake the damaged spectrum. The verification performance was evaluated through a speaker verification experiment with signals corrupted by white noise under different signal to noise ratios. The results were promising since they reflected a relevant increase of speaker verification performance, applying MF approach with this combination of techniques.

1 Introduction

Dealing with noisy signals is a fact in real life, background noise can markedly degrade performance of any speaker verification system. In order to handle environmental noise to improve the robustness of verification performance, many techniques have been proposed [1]. Most of them were originally designed and applied in speech verification application. MF method [2] is an example of that.

MF approach is a group of techniques developed to compensate for noise. Unlike other compensation methods MF does not require to know a priori the characteristics of noise to handle unknown noise. Because of that, it has a lot of potential to ensure robustness in speaker verification applications which process speech signals acquired in noisy environmental conditions with unknown features. This situation is very frequent in real applications.

The MF approach has two steps. The first determines the level of noise corruption in each time-frequency region of speech spectrum to set up a map of binary labels called spectrographic mask. The mask tags as unreliable *(U)* the time-frequency spectral components that are so corrupted by noise that can cause poor verification performance, and tags as reliable *(R)* the time-frequency spectral components that are not very corrupted by noise. The second step is compensation of unreliable region, it could be bypassing the spectral unreliable locations in the verification process, known as marginalization, or reconstructing

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 220–227, 2010.

unreliable spectrum location and keeping the verification process with the new reconstructed spectrum.

Until now, most of the MF development has occurred on the speech verification field, while only a few works have been done on speaker verification [3][4][5][6]. This work presents a novel combination of MF techniques for robust speaker verification with noisy speech. To estimate the MF mask we proposed the use of SNR criterion. For MF compensation we proposed to use a reconstruction method which estimates U components from R ones. This kind of reconstruction has not been previously used for speaker verification. We evaluate the performance impact of this MF setup through speaker verification experiment in noisy environments.

From now on, this paper is organized as follows. Section 2 describes mask estimation technique. Section 3 explains the MF compensation technique used. Section 4 presents speaker verification experiments and results. Finally, section 5 a discussion of results and conclusions.

2 Mask Estimation

The success of the MF approach in providing robustness to speaker verification system will depend on the mask accuracy [2]. To estimate the masks, the SNR criterion is the most widely used in previous works because of SNR-based masks are very easy to compute [7].

In this paper we proposed, as MF detector, the identification of U spectral components based on spectral enhancement technique used frequently in speech processing. This approach was applied to MF mask estimation in the previous work [8]. This is an effective technique in the detection of corrupted components that is known as Negative Energy Criterion.

This method uses a frame by frame spectral subtraction algorithm as MF detector and is based on an estimated noise spectrum. The reliability decision of spectral components is done following this rule:

$$\begin{array}{lll} |Y(f,s)|^2 \leq |\hat{N}(f,s)|^2 & then & Y(f,s) \leftarrow U \\ |Y(f,s)|^2 > |\hat{N}(f,s)|^2 & then & Y(f,s) \leftarrow R \end{array} \tag{1}$$

where, f and s are the frame (time) and subband (frequency) spectrographic representation of the signal power spectrum, respectively. If the power spectrum in a component is less than the estimated noise power spectrum in it, this component is assumed as U, otherwise the component is tagged as R.

3 Cluster-Based Reconstruction

Until now, most speaker verification systems using the MF approach, to improve performance in noisy environments, have been based on modifying the classifier to work with the reliable components of the spectrographic representation of the speech signal. That is the case of the works of Drygajlo et al. [8] or Padilla et al.

[3]. In these systems, the unreliable log-Mel spectral components are integrated out of the GMM distributions to get the speaker likelihood. This technique is known as marginalization.

Marginalization has several drawbacks. On the one hand, recognizers are constrained to use Mel spectral features that are known to produce worse performance than Mel frequency cepstral coefficients (MFCC). On the other side, by using incomplete spectrographic data we are not able to apply certain feature processing steps that are known to improve considerably the results. These processing steps include mean normalization, feature warping [9] or added time derivatives .

For these reasons, in this paper we are taking an alternative approach by trying to estimate the true values of the unreliable spectrographic components from the reliable ones. Once we get the complete time frequency representation of the signal, we are able to compute MFCC features, and apply whatever post-processing step to the features. Besides, we do not need to modify the recognizer so we can use anyone at our disposal. The algorithm we have chosen to compensate for the U components is cluster-based reconstruction which has proven to be very effective in speech verification tasks as it is reported in the work of Raj et al. [10] [11].

3.1 The Algorithm

The Cluster-based Reconstruction (CBR) algorithm estimates the U components of the spectral vector from the R ones of the same vector using a statistical model that relates both of them. This method is based on the assumption that the sequence of observations is an independent, identically distributed random process. This assumption is used by the most successful text independent speaker verification approaches too. Therefore, it is expected to have good results for MF compensation in speaker verification systems.

This algorithm models the distribution of log-Mel spectral vectors for clean signals as a mixture of Gaussian distributed clusters. The mean, covariance and a priori probability of each cluster can be estimated from a training corpus using maximum likelihood estimation via the expectation maximization (EM) algorithm [12].

Let Y be the noisy spectral vector and X the reconstructed spectral vector and let Y_r, X_r and Y_u, X_u be their R and U components respectively. The first step to compensate for the U components is to determine the noisy vector probability of belonging to each cluster. This is given by

$$P(k|Y) = \frac{w_k P(Y|k)}{\sum_{j=1}^{k} w_j P(Y|j)} \tag{2}$$

where w_k is the a priory cluster probability.

To calculate the term $P(Y|k)$ we have to take into account that Y has R and U components, and that $X_r = Y_r$ and $X_u \leq Y_u$ for additive noises. Therefore we can evaluate the Gaussian distribution in the R components and integrate out the U ones. This integration supposes additive noise so, the estimated U components need to be less than the measured components

$$P(Y|k) = P(X_r, X_u \leq Y_u|k) = \int_{-\infty}^{Y_u} P(X_r, X_u|k) dX_u \tag{3}$$

If we suppose that the covariance matrices are diagonal this can be written as

$$\begin{gathered} P(Y|k) = \mathit{\Pi}_{i|X_i \epsilon X_r} \tfrac{1}{\sqrt{2\pi}\sigma_{ki}} exp(-\tfrac{1}{2}\tfrac{(X_i-\mu_{ki})^2}{\sigma_{ki}^2}) \times \\ \mathit{\Pi}_{i|X_i \epsilon X_u} \tfrac{1}{2}(1 + erf(\tfrac{Y_i-\mu_{ki}}{\sqrt{2}\sigma_{ki}})) \end{gathered} \tag{4}$$

where erf is the Gauss error function.

We can get an estimation of the clean value of the unreliable components from each cluster based on its distribution maximizing its likelihood given the measured reliable and unreliable components as

$$\hat{X}_u^k = \arg\max_{X_u}\{P(X_u|k, X_u \leq Y_u, X_r = Y_r)\} \tag{5}$$

Assuming diagonal covariance matrices this can be reduced to

$$\hat{X}_u^k = min(Y_u, \mu_{kr}) \tag{6}$$

where μ_{kr} is the Gaussian means of the unreliable components of the associated cluster.

Finally, we can get the overall unreliable components using the posterior membership probabilities to combine, by a weighted sum, the unreliable components estimations given by each cluster.

$$\hat{X}_u = \sum_{k=1}^{K} P(k|Y)\hat{X}_u^k \tag{7}$$

Once we have recovered the full Mel spectral vector, we are able to calculate the MFCC with their time derivatives and apply any preprocessing technique we need prior to the recognizer input.

4 Experiments and Results

In order to evaluate the behavior of the MFs techniques combination in front of corrupted signals, a speaker verification experiment was carried out using the 1conv4w-1conv4w task of the 2006 NIST SRE [13].

4.1 Detection and Compensation of Unreliable Components

To implement the mask estimator based on spectral subtraction we used the classical algorithm of Berouti et al. [14] and the noise estimator of Martin work [15].

The noisy signals were segmented with 25 msec. Hamming window overlapped 15 msec. and passed through 24 Mel filters bank. Then, noise estimator was applied, taking decision of reliability presented in equation 1, to obtain the unreliable components of the noise corrupted speech.

Once the mask estimation was done, the Cluster-based Reconstruction algorithm makes an estimation of the unreliable components. These reconstructed log-Mel spectra are then used to calculate the MFCC features that will be the input to the speaker verification system.

4.2 Speaker Verification Protocol

In this task, the enrollment and test utterances contain around 2 minutes of speech after voice activity detection. There are a total of 810 target models with 3176 true trials and 42079 false trials. It has used clean speech to train the target models and contaminated test signals with different levels of white noise selected to get several mean SNR, from 5 to 20 dB.

Our acoustic features are 15 MFCC plus first and second derivatives and C0 derivatives resulting in a total of 47 features. On the one hand, we have got results using no feature normalization at all to prove the capacity of our MF approach to cope with noise on its own. On the other hand, we have repeated the experiments using feature warping over 3 seconds in order to proof the benefits of being able to use feature normalization techniques together with the MF approach.

A gender dependent Universal Background Model (UBM) of 512 Gaussians is used. This model is trained using NIST SRE 2004 database containing 124 male and 184 female speakers with several utterances each one of them. The means of target models are adapted from the UBM using relevance MAP [16]. Classification is performed evaluating the log-likelihood ratio between the target and the UBM model for the test signal. Gender dependent cluster models for CBR are trained from the same dataset as UBM using different number of Gaussians.

4.3 Results

The first experiment we have conducted was intended to determine the optimal number of Gaussians needed for reconstruction. For that purpose, we have got results comparing baseline and MF cluster-based reconstruction with different number of clusters between 64 and 1024 using test signals contaminated with a SNR of 10 dB. The experiment has been repeated using feature warping and no feature normalization. In Table 1, we show the equal error rate (EER) and improvement percentage relative to the baseline of this experiment.

Figure 1 shows DET curves using no feature normalization and feature warping respectively, results with number of cluster over 256 are not plotted in order to preserve clarity.

We have got an amazing improvement when no feature normalization is applied nearly reaching clean signal performance. When using feature warping the challenge is bigger, but MF achieves a considerable improvement. The great capacity of feature warping of increasing robustness against channel mismatch,

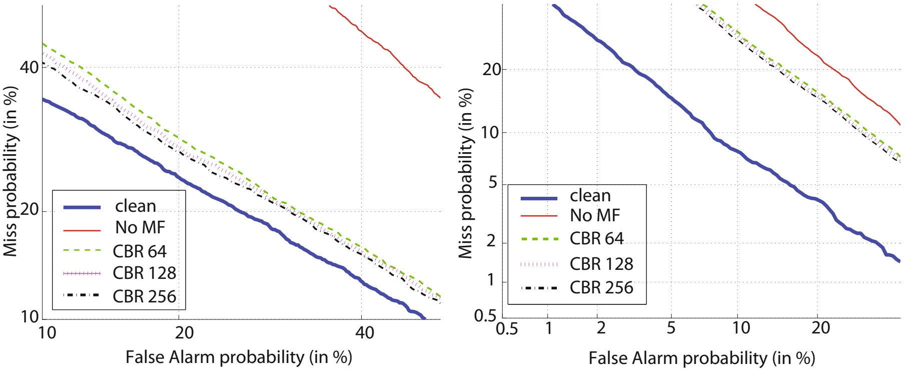

Fig. 1. DET curves to SNR=10 dB (left), with features normalization (right)

Table 1. EER and Improvements to SNR = 10 dB

	No Feat. Norm.		Feat. Warp.	
	EER(%)	Δ(%)	EER(%)	Δ(%)
clean	22.3		8.7	
No MF	42.9	0	21	0
CBR 64	25.2	41.2	17.7	15.4
CBR 128	24.7	42.4	17.4	17.1
CBR 256	24.2	43.6	17.1	18.6
CBR 512	24.9	41.9	16.8	20
CBR 1024	24.3	43.5	17.3	17.6

Table 2. EER and Improvements to SNR = 5-20 dB

SNR(dB)	20	15	10	5
EER(%) No MF	29.8	36.9	42.9	46.8
EER(%) MF	21.8	22.5	24.2	29.5
Δ(%)	26.8	39	43.6	36.9
Feature Norm.				
EER(%) No MF	13.37	16.95	21	27.2
EER(%) MF	11.5	13.5	17.1	22.5
Δ(%)	14.5	20.3	18.6	17.28

additive noise or even headset non-linearity it is well known. As a matter of fact, most sites participating in NIST evaluations use it in their systems. As we can see in Table 1, feature warping on its own is able to provide better results than MF compensation alone. That means it does a great deal of the same job as MF does. However, the benefits of using both techniques together are not negligible producing around a 17 percent of improvement compared to using feature warping only. This encourages us to think reconstruction of missing spectral component is the right path to follow in order to take advantage easily of the existing techniques to build robust speaker verification systems.

Results show there is little improvement as we increase the number of clusters getting the best performance with 256 with no feature normalization and 512 with feature warping. We have found there is no improvement if we use more clusters. This could be explained by the fact that if we increase the number of clusters, they become more similar among them. Considering that cluster membership is estimated using only the reliable components of the spectrogram, it becomes more difficult to select precisely the best cluster as the number of clusters rises.

We have repeated the experiment using signals contaminated with SNR between 5 and 20 dB. This time we have used only 256 clusters, what seems a good choice given the previous results. In Table 2, we give a summary of the obtained results. We have got interesting improvements for all SNR tested. Something curious we note is that with no feature normalization and a SNR of 20 dB EER outperforms clean signal one. We expected a more important decrease of the improvement with low SNR due to the fact that we have less reliable components to make the spectral reconstruction but results are quite good.

5 Conclusions and Future Work

In this paper the proposed MF techniques combination has shown its potentiality in providing robustness for speaker verification systems. The results obtained with MF alone or in combination with feature normalization produced an important increase of verification performance. It is convenient to highlight some ideas:

Improvement obtained in speaker verification results show that SNR criterion is an effective method when trying to obtain the reliability of the corrupted speech spectral components. However the enhancement of SNR contributes to increase speech quality, but does not necessarily ensure the improvement of verification performance, so in the future we will focus on criteria that use representative speaker features. We will evaluate mask estimation methods based on spectral features classification such as Seltzer et. al work [11].

Since mask estimation is the prior step in MF approach, we do not lose sight of the MF compensation step. In this work we have used a reconstruction technique originally designed for speech verification. We must take into account the fact that we have used speaker independent cluster models. This means that reconstructed features will be made more speaker independent too. In speaker verification applications this is a great drawback. Despite that, results show improvements since noise compensation is more important than the effect of using speaker independent models. Nevertheless, we think we could get even better results using cluster models adapted to the test signal. Future work will be oriented in that direction.

On the other side, we must take into account the fact that GMM distributions with diagonal covariance matrices have limited correlation information between features. In future work, we plan to perform MF reconstruction using more complex distributions that should be able to perform a more precise estimation of the U components values. Examples of these models are GMM with full covariance matrices or graphical models [17]. Graphical models have the capacity of modeling correlations between features or groups of features at any level of complexity, what can be very promising for the MF approach.

Acknowledgements

This work has been supported by the Spanish Government through national project TIN2008-06856-C0504.

References

[1] Benesty, J., Chen, J., Huang, Y., Cohen, I.: Noise Reduction in Speech Processing. Springer Topics in Signal Processing 2 (2009)

[2] Raj, B., Stern, R.: Missing-Feature Approaches in Speech Recognition. In: IEEE Signal Proc. Magazine (2005)

[3] Padilla, M., Quatieri, T., Reynolds, D.: MF Theory with Soft Spectral Subtraction for Speaker Verification (2006)

[4] Ming, J., Hazen, T., Glass, J.R., Reynolds, D.A.: Robust Speaker Recognition in Noisy Conditions. IEEE Trans. on Speech and Audio Proc. 15, 1711–1723 (2007)

[5] Pullella, D., Kuhne, M., Togneri, R.: Robust Speaker Identification Using Combined Feature Selection and Missing Data Recognition. In: ICASSP (2008)

[6] Kuhne, M., Pullella, D., Togneri, R., Nordholm, S.: Towards the use of full covariance models for missing data speaker recognition. In: ICASSP (2008)

[7] Cerisara, C., Demange, S., Haton, J.-P.: On noise masking for automatic missing data speech recognition: a survey and discussion. Computer Speech and Language 21(3), 443–457 (2007)

[8] Drygajlo, A., El-Maliki, M.: Speaker Verification in Noisy Enviroments with Combined Spectral Subtraction and MF Theory. In: Signal Proc. Laboratory, Swiss Federal Institute of Technology at Lausanne (1998)

[9] Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. Speaker Odyssey (2001)

[10] Raj, B., Seltzer, M., Stern, R.M.: Reconstruction of MFs for robust speech recognition. Speech Communication 43 (2004)

[11] Seltzer, M., Raj, B., Stern, R.M.: A Bayesian classifier for spectrographic mask estimation for MF speech recognition. Speech Communication 43 (2004)

[12] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (1977)

[13] The NIST year, speaker recognition evaluation plan (2006), http://www.nist.gov/speech/tests/spk/2006/index.htm

[14] Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: IEEE ICASSP (1979)

[15] Martin, R.: Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. IEEE Trans. on Speech and Audio Proc. 9 (2001)

[16] Reynolds, D., Quatieri, T., Dunn, R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Proc. 10 (2000)

[17] Bilmes, J.: Graphical Models and Automatic Speech Recognition. Mathematical Foundations of Speech and Language Proc., 191–235 (2004)