SPEAKER VERIFICATION IN NOISY ENVIRONEMENTS WITH COMBINED SPECTRAL SUBTRACTION AND MISSING FEATURE THEORY

Andrzej Drygajlo and Mounir El-Maliki

Signal Processing Laboratory
Swiss Federal Institute of Technology at Lausanne
CH-1015 Lausanne, Swizerland
e-mail: {Andrzej.Drygajlo,Mounir.Elmaliki}@epfl.ch

ABSTRACT

In the framework of Gaussian mixture models (GMMs) [1], we present a new approach towards robust automatic speaker verification (SV) in adverse conditions. This new and simple approach is based on the combination of a speech enhancement using traditional spectral subtraction, and a missing feature compensation to dynamically modify the probability computations performed in GMM recognizers. The identity of spectral features missing due to noise masking is provided by the spectral subtraction algorithm. Previous works have demonstrated that the missing feature modeling method succeeds in speech recognition with some artificially generated interruptions, filtering and noises [2, 3]. In this paper, we show that this method also improves noise compensation techniques used for speaker verification in more realistic conditions.

1. INTRODUCTION

The ability to deal with missing and noisy features is vital in automatic speaker recognition over the telephone lines. It is well-known that speech degraded by background noise renders the performance of many realistic speaker recognition systems unacceptable. If a noise compensation algorithm which sufficiently reduces the effects of background noise could be derived, then existing GMM-based speaker recognition techniques, formulated in noise-free settings, could be employed in noisy environments. In order to improve the recognition performance in very noisy conditions, the enhancement techniques are needed.

In this work, we study how to adapt clean speech models for a signal enhanced by the spectral subtraction (SS) method. The classical SS schemes improve the signal-to-noise ratio (SNR) but at the expense of a signal distortion. In automatic speaker verification, there is no need to reconstruct the speech signal. The performance measure, as it is given by the equal error rate (EER), is simplified compared to the speech enhancement procedure. For a system whose aim is to decrease the EER, it is important to take into account some properties of the human auditory system. The auditory representations of clean speech contain much redundancy. Arguably, it is this redundancy which enables listeners to recognize speaker in adverse conditions. Under the assumption that some time-frequency regions are too heavily masked to derive any valuable data, the auditory system faces the missing data problem. In automatic speaker recognition terms, we face the missing features problem.

This paper describes our recent attempts to adapt dynamically the stochastic automatic speaker verification framework of GMMs to handle the missing features problem with the help of traditional spectral subtraction method. In this case, the traditional spectral subtraction algorithm is used as a simple missing feature detector, and not as a pure enhancement system. Recognition results are reported for various types of noise, tested on a challenging text-independent telephone-quality speaker verification task.

2. MISSING FEATURES IN GAUSSIAN MIXTURE MODELING

The missing feature theory was succesfully applied to a large class of learning algorithms, including feedforward networks [4], normalized radial basis function neural networks [5] and CDHMM [6]. In this paper, Gaussian mixture model is applied, in combination with missing feature theory, for the specific task of speaker verification. GMM models the probability density function (pdf) of the observed spectral features generated from a speech signal by a multi-variate Gaussian mixture density. The GMM pdf is defined as follows:

$$p(\overrightarrow{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i \Phi_i(\overrightarrow{\mathbf{x}}, \mu_i, \Sigma_i) \qquad (1)$$

where $\overrightarrow{\mathbf{x}}$ is a D-dimensional feature vector, M represents the number of uni-modal Gaussian densities Φ_i, each defined by a mean vector μ_i, and a covariance matrix Σ_i and weighted by the mixing proportions p_i.

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \;\; i = 1, \cdots, M$$

represents the set of parameters of the speaker model.

In the case of a diagonal covariance matrix, equation (1) can also be expressed as a combination of the products of uni-variate Gaussian densities:

$$p(\overrightarrow{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j=1}^{D} \Phi_i(x_j, m_{ji}, \sigma_{ji}^2) \tag{2}$$

where m_{ji} is the mean and σ_{ji}^2 the variance of GMM univariate pdf. The parameters of the speaker model λ are estimated during training using clean speech by the Expectation-Maximization (EM) algorithm [7]. This allows to consider the whole spectral components of feature vectors, in the absence of a noisy source, as useful information for training, and then, no modification of the conventional training procedure will be done. On the other hand, for the recognition process, if we assume that speech samples are corrupted by a masking noise or affected by filtering or interruptions, the feature vector $\overrightarrow{\mathbf{x}} = (\overrightarrow{\mathbf{x_p}}\overrightarrow{\mathbf{x_m}})$, will be composed by two subvectors $\overrightarrow{\mathbf{x_p}}$ and $\overrightarrow{\mathbf{x_m}}$ representing present and missing components, respectively. As a result, equation (2) takes the following structure:

$$p(\overrightarrow{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j}^{pr} \Phi_i(x_j, m_{ji}, \sigma_{ji}^2) \prod_{j}^{mi} \Phi_i(x_j, m_{ji}, \sigma_{ji}^2) \tag{3}$$

where pr denotes the present feature and mi the missing one. Missing feature compensation eliminates the contribution of missing data from the computation of the GMM pdf. In this case, the modified pdf, computed only on partial data, preserves a mixed Gaussian form [6], as expressed in equation (4):

$$p(\overrightarrow{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j}^{present} \Phi_i(x_j, m_{ji}, \sigma_{ji}^2) \tag{4}$$

3. SPECTRAL SUBTRACTION AND MISSING FEATURES

Let $y(n)$ be the speech samples affected by an additive stationary noise $n(n)$:

$$y(n) = s(n) + n(n) \tag{5}$$

A simple technique for increasing the robustness of a speaker recognition system is to apply an enhancement system as a pre-processing stage to remove the effect of the estimated noise from each spectral component. This is done by using the power spectral subtraction algorithm. The short-time power spectrum of the enhanced speech is based on the following noise reduction rule:

$$|\hat{S}_m(\omega)|^2 = \begin{cases} |Y_m(\omega)|^2 - |\bar{N}(\omega)|^2 & \text{if } |Y_m(\omega)|^2 > |\bar{N}(\omega)|^2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $|Y_m(\omega)|^2$ is the power spectrum of the current noisy speech frame, and $|\bar{N}(\omega)|^2$ is the averaged noise power estimate. Once the subtraction has been computed in the spectral domain with equation (6), the enhanced speech signal is obtained with the next relationship:

$$\hat{s}(n) = IFFT[|\hat{S}(\omega)|.e^{(jargY(\omega))}] \tag{7}$$

As any subtractive-type algorithm, spectral subtraction introduces a residual noise after the enhancement process, given by:

$$r(n) = s(n) - \hat{s}(n) \tag{8}$$

The residual noise $r(n)$ has a musical nature and is completely different from the original noise. $r(n)$ can sometimes be more disturbing not only for human listeners, but also for a speaker recognition system. This is due to the presence of tones at random frequencies. The existence of these tones is caused by the null term in equation (6) [8]. Hence, according to sections 1 and 2, the combination of missing feature theory with spectral subtraction is motivated by three main reasons:

- Spectral subtraction algorithm is a simple missing feature detector performing on a frame-by-frame basis. Indeed, when $|Y_m(\omega)|^2 < |\hat{N}(\omega)|^2$, the power spectrum component $|Y_m(\omega)|^2$ at frame m, is considered inappropriate for replacement by any estimate. It becomes a missing feature and can be ignored in the calculation of the GMM pdf.

- The valuable partial data is not only detected by spectral subtraction algorithm, but also enhanced according to equation (6).

- Since there is no need to reconstruct the enhanced signal, and the classification uses only the valuable data, the influence of residual musical noise on recognition accuracy is attenuated.

The approach of combination of the two techniques is depicted in Fig. 1. First, the averaged noise estimate $|\bar{N}(\omega)|^2$ is calculated with the help of a speech/pause detector. Then, the spectral subtraction algorithm is performed using the short-power spectrum of the corrupted speech and the noise estimate. In order to reduce the number of components of feature vectors, log-energies of the Bark filter bank are used [9]. They are computed as follows:

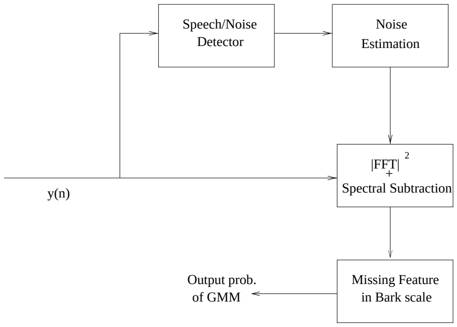
Speech/Noise Detector
Noise Estimation
$|FFT|^2$ + Spectral Subtraction
y(n)
Missing Feature in Bark scale
Output prob. of GMM

Figure 1: *Block diagram of the proposed system.*

$$\hat{S}_{Bark_i} = log \sum_{j} |\hat{S}(\omega_j)|^2 \quad i = 1, \cdots, 14 \tag{9}$$

where $Bark_i$ represents the ith position of Bark band, and $\frac{\omega_j}{2\pi} \in Bark_i$. A new selection of missing features in this filter bank is done using the following rule:

$$if \ \ |\hat{S}\omega_j)|^2 = 0, \ \frac{\omega_j}{2\pi} \in Bark_i \quad then \ \ \hat{S}_{Bark_i} = 0 \tag{10}$$

4. EXPERIMENTAL RESULTS

A set of 22 speakers, 13 males and 9 females, with the same dialect region ’dr1’ was selected from the NTIMIT corpus [10]. All the speakers are presented by 10 sentences, and each sentence lasts, in the average, about 3 seconds. Gaussian mixture models are built using 8 sentences for each speaker. The speech data is processed by a silence removing algorithm and a 32 ms Hanning window is applied to the speech samples with 50% of overlap. The spectral subtraction is performed for speech enhancement solely using the power spectrum of each frame. Finally, the last two sentences which are not included to build speaker model are used for the verification experiments. A total of 44 genuine accesses and 924 impostors attempts is performed for the evaluation.

4.1. Recognition accuracy with additive artificial colored noise and aircraft-cockpit noise

The first experiment undertaken to evaluate the performance of our approach is based on adding, to the test speech signals, an artificial colored noise (ACN) concentrated in one of the 14 bands of the Bark filter bank (Fig. 2). The artificial colored noise is obtained in our experiments by a band-passed white Gaussian noise with the cutoff frequencies of 630 Hz and 770 Hz. The SNR level is chosen equal to 9 dB. As seen in table 1, without applying an enhancement technique to the noisy data, the EER reaches 24%, while combined spectral subtraction and missing feature compensation (MFC) decreases the EER with a reduction of about 50% to EER=12.92%. The spectral subtraction algorithm gives an EER=18.85% which is worse than the one provided by the proposed approach.

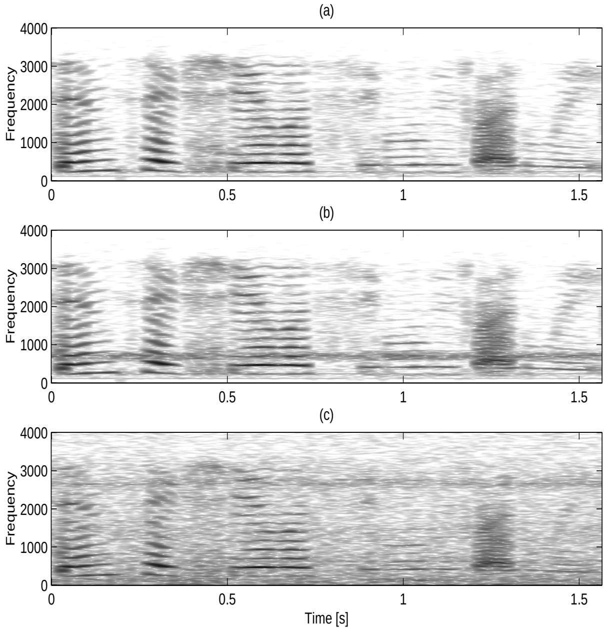

Figure 2: *Spectrograms of: (a) clean speech uttered by a female, (b) speech with additive artificial colored noise, (c) speech with additive F16 cockpit noise*

Experiments	EER[%]	EER[%]
(Noise,SNR)	(ACN,9dB	(F16,9dB)
clean speech	11.5	11.5
noisy speech	24	37.2
SS	18.85	28.2
combined SS & MFC	12.92	20.7

Table 1: *EER in presence of ACN and F16 cockpit noise.*

A second experiment is carried out on corrupted speech by an aircraft cockpit noise. This noise was selected from the NOISEX-92 database. It has a strong energy below 1600 Hz and an important presence at a frequency of 2700 Hz approximately (Fig. 2). Table 1 shows a drastic degradation of the performance of the speaker verification system. However, spectral subtraction succeeds to decrease the EER up to 28.2%. An EER=20.7% is obtained by the missing feature compensation combined with spectral subtraction.

4.2. Experiments with white Gaussian noise

Gaussian white noise, at different SNR levels, is added to the test speech data to assess the robustness of the proposed approach in an application where many spectro-temporal regions of the speech are corrupted and masked by noise. Table 2 gives a summary of the obtained results. The introduced approach gives better results than a simple spectral subtraction performed in a pre-processing stage. A reduction of the EER of about 41% is obtained when SNR=15dB, while this EER is attenuated by a factor of about 28% when SNR=9dB. The difference between the percentage of the EER reduction in the two experiments could be explained by the fact that in a very noisy environement, a significant number of feature components is greatly masked by noise, and then considered as missing data. Thus, the verification is performed only on smaller amount of valuable data.

Experiments (SNR)	EER[%] (9dB)	EER[%] (12dB)	EER[%] (15dB)
noisy speech	43.6	37.5	34.8
SS	38	34.38	31.25
combined SS & MFC	27.1	21.9	18.5

Table 2: *EER in presence of white Gaussian noise.*

5. CONCLUSIONS

This paper presents missing feature modeling combined with spectral subtraction for the speaker verification task in the presence of an additive noise. Several experiments were carried out to evaluate the effectiveness of this combination and the results were compared with those obtained by applying the classical spectral subtraction algorithm as a pre-processing stage for speech enhancement. In all these experiments, the combined approach has significantly improved the recognition rate in comparison with the conventional enhancement technique.

6. REFERENCES

[1] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models", *Speech Communication*, vol. 17, pp. 91–108, 1995.

[2] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition", *in Proc. ICASSP*, pp. 863–866, Munich, April 1997.

[3] R. P. Lippman and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise", *in Proc. EUROSPEECH*, vol. 1, pp. KN 37–40, Rhodes, Sep. 1997.

[4] V. Tresp, R. Neuneier, and S. Ahmed, "Efficient methods for dealing with missing data in supervised learning", in G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 7*. Morgan Kauffman, San Mateo, 1995.

[5] S. Ahmed and V. Tresp, "Some solutions to the missing feature problem in vision", in S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pp. 393–400. Morgan Kauffman, San Mateo, 1993.

[6] M. Cooke, P.D. Green, C. Anderson, and D. Abberley, "Recognition of occluded speech by hidden markov models", *in University of Sheffield, Department of Computer Science*. Technical Report TR-94-05-01, 1994.

[7] A. Demspter, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm", *J. Royal Statistical Soc.*, vol. 39, 1977.

[8] J.R Deller, J.G. Proakis a, and J.H.L Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, 1993.

[9] E. Zwicker and E. Terhardt, "Analytical expressions for critical band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. America*, vol. 68, pp. 1523–1525, Dec. 1980.

[10] C. Jankowsky, A. Kalyanswamy, S. Basson, and J. Spitz, "Ntimit: a phonetically balanced, continous speech telephone bandwidth speech database: Specifications and status", *in Proc. ICASSP*, vol. 1, p. 109, Albuquerque, 3-6 Avr. 1990.