SPECTRAL SUBTRACTION AND MISSING FEATURE MODELING FOR SPEAKER VERIFICATION

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

ABSTRACT

This paper addresses the problem of robust text-independent speaker verification when some of the features for the target signal are heavily masked by noise. In the framework of Gaussian mixture models (GMMs), a new approach based on the spectral subtraction technique and the statistical missing feature compensation is presented. The identity of spectral features missing due to noise masking is provided by the spectral subtraction algorithm. Consequently, the statistical missing feature compensation dynamically modifies the probability computations performed in GMM recognizers. The proposed algorithm uses a variation of the generalized spectral subtraction and incorporates in it a criterion based on masking properties of the human auditory system. The originality of the algorithm resides in the fact that instead of using fixed parameters for the noise reduction and missing feature compensation, the noise masking threshold is used to control the enhancement and model compensation processes adaptively, frame-by-frame, hence helping to find the best tradeoff.

1 INTRODUCTION

The problem of enhancing speech degraded by noise for automatic speaker verification over the telephone lines remains largely open, even though many significant techniques have been introduced over the past decades [1]. 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 [2]. In order to improve recognition performance in very noisy conditions, enhancement techniques are needed.

Speech enhancement techniques are mainly applied as a preprocessing stage to many automatic speech and speaker recognition systems [3]. In this paper, we propose a new paradigm for robust speaker verification based on the missing data theory and the generalized spectral subtraction speech enhancement technique.We study how to adapt clean speech models for a signal enhanced by the spectral subtraction (SS) method [3, 4]. The classical SS schemes improve the signal-to-noise ratio (SNR), but at the expense of signal distortion. If both signal distortion and residual noise are minimised, then a signal with better features and lower variability is obtained. Furthermore, if these resulting features are to be exploited in a speaker verification system, then the speaker models need to be adapted to the distorted signal.

In automatic speaker recognition 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 speech enhancement. It is important for a system whose aim is to decrease the EER to take into account some properties of the human auditory system. Auditory representations of clean speech contain much redundancy. Arguably, it is this redundancy which enables listeners to recognize speakers in adverse conditions. It is also well known that locally weaker sound components do not contribute to the neuronal output: they are masked and therefore can be considered missing for the purposes of further processing. 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 statistical automatic speaker recognition framework of GMMs to handle the missing features problem with the help of the generalized spectral subtraction method [5, 6, 7]. Missing feature components can either be estimated or ignored by the spectral subtraction technique. In our case, the generalized spectral subtraction algorithm is used as a perceptually tuned missing feature detector with noise reduction, and not as a pure enhancement system. The noise masking properties are modeled by calculating a noise masking threshold. This threshold helps to minimise both residual noise and signal distortion by modifying each frame of speech observation according to the optimum coefficients α and β which are also used to detect the missing features. α is an over-subtraction factor, and β is the

spectral flooring parameter. Recognition results are reported for various types of noise, tested on a challenging text-independent telephone-quality speaker verification task.

2 MISSING FEATURES MODELING IN GMMs

The speech samples could be corrupted with channel noise or/and background noise. Consequently, some time-frequency regions of speech signal are masked by these noises and the missing features problem appears in the observation vectors X. In such a case, each feature vector from the sequence $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ extracted from a speaker utterance consists of two sub-vectors $\mathbf{x}_p$ and $\mathbf{x}_m$. The sub-vectors $\mathbf{x}_p$ and $\mathbf{x}_m$ represent, respectively, the present and the missing feature components.

There are two approaches to deal with missing features [8]. The first approach consists in estimating the missing features. One simple technique, called mean imputation, replaces the values of missing features by means ($\mathbf{x}_m = \mu_m$). Other techniques give an estimation using conditional means and conditional variances given present features.

The second approach ignores the missing features in the likelihood calculation instead of estimating them. The modified likelihood is computed on the basis of the present features $\mathbf{x}_p$ using marginal distribution obtained by integrating the full (original) likelihood function over the missing features. The use of the second approach is motivated by the fact that estimating missing features is often inappropriate. In this paper, the missing features are ignored by the Gaussian mixture models (GMMs) during the recognition phase.

The parametric modeling capabilities of the GMM allow it to model any arbitrarily shaped probability density function (pdf) with a weighted sum of M component Gaussian densities as given by the equation [2]:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \cdot b_i(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is a D-dimensional feature vector, $b_i(\mathbf{x})$ are the component densities and p_i are the mixture weights, $i = 1, \ldots, M$. Each component density is a D-variate Gaussian pdf with mean vector μ_i and covariance matrix Σ_i. The parameters of speaker model λ_s corresponding to the complete Gaussian mixture density are denoted as

$$\lambda_s = (p_i, \mu_i, \Sigma_i) \qquad i = 1, \ldots, M \tag{2}$$

In the presence of noisy data and with diagonal covariance matrix, the GMM pdf takes the following form:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \prod_{\substack{j \\ present}} b_i(x_j, \mu_{ji}, \sigma_{ji}^2) \cdot \prod_{\substack{j \\ missing}} b_i(x_j, \mu_{ji}, \sigma_{ji}^2) \tag{3}$$

where μ_{ji} is the mean and σ_{ji}^2 is the variance of the feature vector component x_j. The modified GMM pdf computed for the partial data $\mathbf{x}_p$ is extracted from the full multivariate Gaussian densities by integrating the full likelihood function $p(\mathbf{x}|\lambda)$ over the missing features $\mathbf{x}_m$. This is equivalent to supressing the second term of the product present in Eq. 3 [9],

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

Although ignoring the missing features could also be used with a full covariance matrix, the diagonal covariance was prefered in the experiments to decrease computation load due to inversing the covariance matrix of present features at each frame.

3 MISSING FEATURE DETECTION BY SPECTRAL SUBTRACTION

The spectral subtraction algorithm was introduced to reduce the spectral effects of acoustically added noise in speech [4]. One of the various implementations of the spectral subtraction algorithm is referred to as generalized power spectrum subtraction [4, 10]. It is expressed as follows:

$$D_m(\omega) = |Y_m(\omega)|^2 - \alpha |\bar{N}(\omega)|^2 \tag{5}$$

and

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

where $|Y_m(\omega)|^2$ is the power spectrum of noisy speech for frame m and $|\bar{N}(\omega)|^2$ represents the averaged power spectrum estimate of noise updated during speech pauses, $\alpha \geq 1$ and $0 < \beta \leq 1$.

This approach is justified in situations where reconstruction of the enhanced signal is required. In this paper, the use of the generalized spectral subtraction procedure is seen not only as a speech enhancement technique but also as an automatic, frame-by-frame, missing feature detector. According to Eq. 6, the short-term energy of the corrupted speech components lying below a threshold proportional to the noise energy is recognized to be deeply affected, and hence, the estimation of the clean speech is unreliable. In the frequency domain, the spectral components below the spectral floor

are unreliable for the classifier and should be ignored. In such a case, they can be classified as missing features and the automatic speaker verification system can drop them from the recognition process. Under the assumption that noise is additive and stationary, the missing feature detection using generalized spectral subtraction can operate in critical subbands.

Ignoring missing data means attempting to classify the feature components solely on the basis of the present information. In this case, the Gaussian Mixture Model (GMM) recognizer should be dynamically modified as presented in Section 2.

Spectral subtraction based enhancement systems are limited by a tradeoff between noise reduction and speech distortion. This tradeoff determines the choice of the parameters α and β in Eqs 5 and 6.

Indeed, at low SNRs, it is impossible to simultaneously minimize speech distortion and residual noise. In our case, we are concerned with reducing noise and increasing EER, while keeping the residual noise and the target signal distortion acceptable to the recognition system. This is done by adapting the prameters α and β in time and frequency based on masking properties.

The originality of the algorithm proposed in this paper resides in the fact that the noise masking threshold is used to adaptively control the enhancement process. Instead of keeping α and β fixed as in [6], the optimal parameters are computed for each frame and whithin each frequency band.

The masking model used in perceptual speech enhancement allows a threshold to be computed, which is then used to control the residual noise and signal distortion. This model takes into account only simultaneous masking (masking in the frequency domain). It shows good performances in speech enhancement, even though it does not take into account temporal masking [11]. Hence, this model is an efficient and simple way of incorporating properties of the auditory model in the enhancement process, without adding a great computational load.

The calculation of the masking threshold is described in [12, 10]. It is based on a rough estimate of the short-time magnitude and it is composed of the steps presented in Fig. 1.

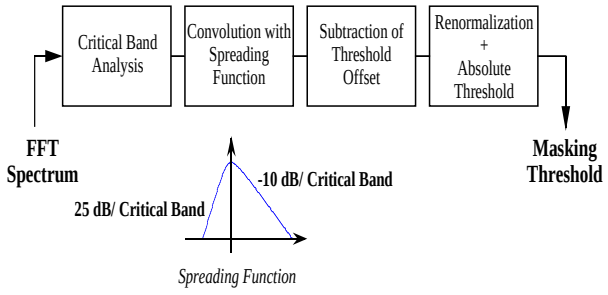

Figure 1: *Calculation of the masking threshold.*

For each frame m, the minimum of the masking threshold $T_m(\omega)$ corresponds to the maxima of the parameters $\alpha_m(\omega)$ and $\beta_m(\omega)$. The adaptation of the subtraction parameters is performed with the following relations: $\alpha_m(\omega) = F[\alpha_{min}, \alpha_{max}, T_m(\omega)]$, $\beta_m(\omega) = F[\beta_{min}, \beta_{max}, T_m(\omega)]$. The minimal and maximal values of α and β determine the tradeoff between residual noise and speech distortion. A number of experiments with different noise types and levels have been performed to select the appropriate values for these parameters. The following values have been chosen in order to obtain a good tradeoff for adaptation of the over-subtraction ($\alpha_{min} = 1.5, \alpha_{max} = 4$) and spectral flooring ($\beta_{min} = 0.001, \beta_{max} = 0.02$). A reduction of α_{max} increases residual noise but reduces speech distortion, while a reduction of β_{max} increases also the residual noise but reduces the background noise remaining in the enhanced speech.

4 EXPERIMENTAL RESULTS

A subset of telephone quality NTIMIT speech corpus composed of 22 speakers was selected for the experiments. This database contains the same speech as TIMIT recorded over local and long distance telephone loops. The GMM classifier is trained on 8 sentences uttered by each of the 22 chosen speakers and the speaker model is built with 32 Gaussian pdfs. Two sentences are used in the test process. Spectral analysis is done by computing the log-energies of 14 auditory critical bands every 16 ms. A 32 ms Hanning window is applied to the speech samples.

An artificial white Gaussian noise and a real F16 airplane cockpit noise selected from the NOISEX-92 database were added to the test data to simulate noisy environments. Several experiments were undertaken with different SNRs:

- A: speaker verification using GMMs with no noise compensation techniques.
- B: speaker verification using GMMs with the generalized spectral subtraction technique (GSS) ($\alpha = 3$ and $\beta = 0.001$) as a speech enhancement method in a pre-processing stage.
- C: speaker verification using GMMs with the missing features compensation and generalized spectral subtraction technique. The factor $\alpha = 3$ gave the best scores during preliminary experiments [6].
- D: speaker verification using GMMs with the missing features compensation and generalized spectral subtraction method based on masking properties of the human auditory system.

Figs 2 and 3 present the evolution of equal error rate versus signal-to-noise ratio. At low SNRs, the generalized spectral subtraction technique improves the recognition rate. In all experiments, the combination of GSS

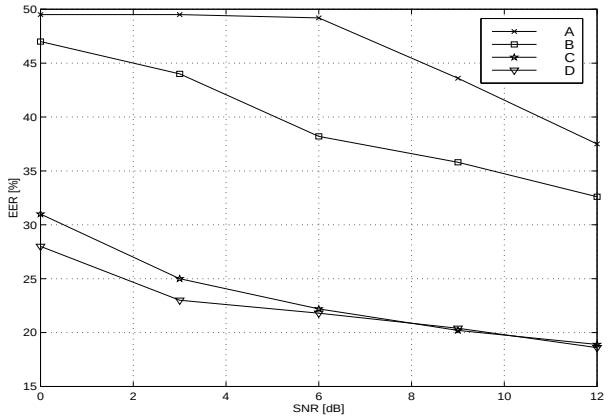

Figure 2: EER *in the presence of white Gaussian noise*

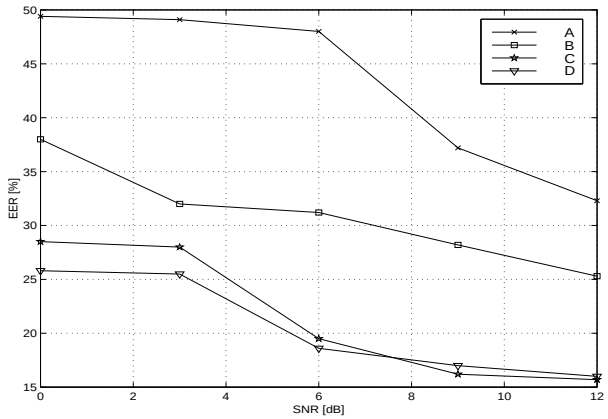

Figure 3: EER *in the presence of airplane cockpit noise.*

and missing feature compensation decreases the EER and performs better than the classical GSS technique in pre-processing stage. Adaptive speech enhancement and missing feature detection based on masking threshold allow more flexibility in the selection of optimal parameters. Therefore, missing feature compensation with dynamic adaptation of GSS parameters outperforms the simple GSS with fixed parameters (α and β), especially when SNR < 6 dB.

5 CONCLUSIONS

According to the spectro-temporal characteristics of the additive noise and the speech signal, the missing features could vary in the time and in the frequency ranges. The use of the generalized spectral subtraction method as an automatic missing feature detector is attractive, as it provides the classifier with a prior knowledge of the missing data in a dynamic way. The main criticism of missing data techniques is that they assume that one already knows what data is missing. On the other hand, the spectral subtraction techniques are too inaccurate to give satisfactory speech enhancement, particularly in heavily masked regions. In combining these two techniques, the strong points from each method are used to full advantage, while the weak points are overcome.

References

[1] J. Ortega-García and J. Gonzàlez-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition", *in Proc. ICSLP*, vol. 2, pp. 929–932, Oct. 1996.

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

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *in Proc. ICASSP'79*, pp. 208–211, April 1979.

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[5] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory", *in Proc. of ICASSP'98*. To be published, May 1998.

[6] A. Drygajlo and M. El-Maliki, "Use of generalized spectral subtraction and missing feature compensation for robust speaker verification", *in Proc. of RLA2C*, pp. 80–83, Avignon, 1998.

[7] M. El-Maliki and A. Drygajlo, "Statistical modeling and missing feature compensation for noisy speech in forensic speaker recognition", *in Proc. COST-250 Workshop*, pp. 39–44, Ankara, 1998.

[8] M. Cooke, A. Morris, and P.D. Green, "Recognizing occluded speech", *in Proceedings of the ESCA Tutorial and Research Workshop on the Auditory Basis of Speech Perception*, pp. 297–300. Keele University, 15-19 July, 1996.

[9] 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'97*, vol. 1, pp. KN 37–40, Rhodes, Sep. 1997.

[10] N. Virag, "Speech enhancement based on masking properties of the auditory system", *in ICASSP'95*, pp. 796–799, Detroit, May 1995.

[11] B. Carnero and A. Drygajlo, "Perceptual speech coding using time and frequency masking constraints", *in ICASSP'97*, pp. 1363–1367, Munich, 1997.

[12] J.D. Johnston, "Transform coding of audio signal using perceptual noise criterea", *IEEE J. on Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.