ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MISSING FEATURES DETECTION AND HANDLING FOR ROBUST SPEAKER VERIFICATION

Mounir El-Maliki and Andrzej Drygajlo

Signal Processing Laboratory
Swiss Federal Institute of Technology, Lausanne
CH-1015 Lausanne, Switzerland
e-mail: [Mounir.Elmaliki,Andrzej.Drygajlo]@epfl.ch

10.21437/Eurospeech.1999-238

ABSTRACT

This paper addresses the problem of robust text-independent speaker verification in the presence of missing (masked by noise) features. It presents and assesses several missing feature handling approaches. In these approaches, the speech enhancement and missing feature detection are based on the minimum mean-square error (MMSE) spectral amplitude estimator of Ephraim and Malah [1].

1. INTRODUCTION

Automatic speaker recognition performance suffers from a drastic degradation in real-world applications. The fall off of recognition rate is essentially due to background noise. It is possible to reduce the influence of noise on speech signal using pre-processing speech enhancement techniques. Unfortunately, these techniques (e.g. spectral subtraction) fail to estimate the clean speech in frequency bands masked by noise.

In our recent works, we have shown that if masked speech segments are included in the recognition process, they could heavily contribute to a fall off of recognition rate [2,3]. We consider the masked components as a missing part of speech and do not include them for further processing. This approach leads to the use of marginal distribution in the framework of Gaussian mixture modeling (GMM). The spectral features missing due to noise masking are provided by the spectral subtraction method.

We have shown that the task of detecting missing features and ignoring them improves substantially the recognition rate comparing to classical speech enhancement techniques. In this paper, we describe our attempts to improve the performance of the GMM-based speaker verification system by using a more accurate missing features detection and speech enhancement based on the minimum mean-square error (MMSE) spectral amplitude estimator [1].

The second part of our paper treats the problem of missing features estimation. Estimating missing features rather than ignoring them is a difficult task due to the low reliability of the estimation. However, several approaches of missing feature estimation are explored to assess their reliability in a speaker verification system.

2. MISSING FEATURES PROBLEM

Missing features problem occurs when speech segments are filtered, interrupted or masked by noise. A diagram representing the speaker verification system integrating missing feature theory is shown in Figure 1. In the first stage, critical band analysis for feature extraction is performed with the use of log spectral energies according to the Bark scale. In the second stage, the degraded subbands at each frame are detected and then ignored in the recognition process. The GMM-based system calculates the likelihood score using marginal distribution of only present data as in the following equation:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \prod_{j \in present} \Phi_i(x_j, \mu_{ji}, \sigma_{ji}^2) \qquad (1)$$

where μ_{ji} is the mean and σ_{ji}^2 is the variance of the feature vector component x_j. λ is the speaker model and $p_i, \quad i = 1, \cdots, M$ are the mixture weights corresponding to the mono-variate Gaussian densities $\Phi_i, i = 1, \cdots, M$.

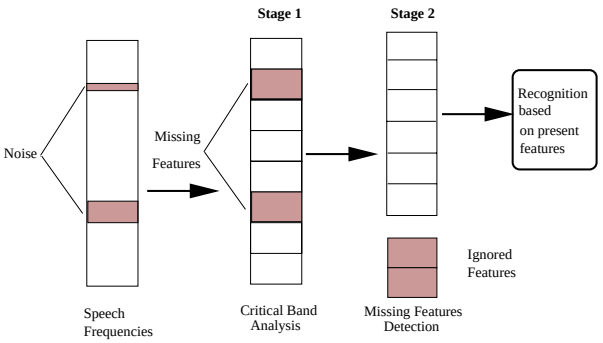

Figure 1: *Block diagram of missing features system.*

3. SPEECH ENHANCEMENT AND MISSING FEATURE DETECTION

The minimum mean-square error (MMSE) spectral amplitude estimation is one of the known applied methods for reducing the influence of additif noise in corrupted speech signal [1]. It can be used in a pre-processing stage for robust speaker verification. The spectral magnitude of the

estimated clean speech is calculated by the MMSE estimator as follows:

$$|\hat{S}_m(\omega)| = G(\omega) \cdot |Y(\omega)| \tag{2}$$

where $|Y(\omega)|$ is the magnitude of noisy speech. The gain function of this estimator is given by the following equation:

$$\begin{aligned} G(\omega) =& \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{\text{SNR}_{post}} \cdot \frac{\text{SNR}_{prio}}{1+\text{SNR}_{prio}}} \cdot \\ & F(\text{SNR}_{post} \cdot \frac{\text{SNR}_{prio}}{1+\text{SNR}_{prio}}). \end{aligned} \tag{3}$$

F represents the following function:

$$F(x) = \exp(\frac{-x}{2})[(1+x) \cdot I_0(\frac{x}{2}) + x \cdot I_1(\frac{x}{2})] \tag{4}$$

where I_0 and I_1 are the modified Bessel functions of zero and first order. The gain function of the MMSE estimator depends on two parameters: *posterior* and *prior* SNRs. The *posterior* SNR is expressed as the ratio of the power spectra of noisy speech $|Y(\omega)|^2$ and the estimated noise during pauses $|\bar{N}(\omega)|^2$.
The *prior* SNR is found to be a dominant parameter and the noise attenuation is strongly dependent on its estimation. It is obtained at the *n*th frame for each sub-band by maximum likelihood estimation approach:

$$\text{SNR}_{prio} = \begin{cases} \overline{\text{SNR}}_{post}(n) - 1 & \text{if } \overline{\text{SNR}}_{post}(n) - 1 \geq 0 \\ 0 & otherwise \end{cases} \tag{5}$$

where:

$$\overline{\text{SNR}}_{post}(n) = \alpha \overline{\text{SNR}}_{post}(n-1) + (1-\alpha) \frac{\text{SNR}_{post}(n)}{\beta} \tag{6}$$

and $0 \leq \alpha < 1$ and $\beta \geq 1$. β is the noise overestimation factor similar to the one used in the spectral subtraction technique [4].
Eq. 5 shows that the maximum likelihood method for the *prior* SNR estimation produces zero values in frequency bands where the estimated noise dominates speech (i.e. $\overline{\text{SNR}}_{post}(n) - 1 \leq 0$). Consequently, the gain function of the MMSE spectral estimator produces zero values as an estimate of the clean speech in these frequency bands. In such a case, these frequency bands components can be classified as missing features since the MMSE could not give a reliable estimation due to noise masking. The automatic system can drop them from the recognition process.

4. MISSING FEATURE HANDLING

The probability density function (pdf) of speaker model is represented as

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i \Phi_i(\mathbf{x}, \mu_i, \Sigma_i) \tag{7}$$

where p_i is the mixture weight and Φ_i is a D-dimensional Gaussian density defined by the mean vector μ_i and the diagonal covariance matrix Σ_i. The background noise can generally be modeled by a mixture of Gaussian pdfs

$$p(\mathbf{x}|\lambda_n) = \sum_{i=1}^{D} b_i^n \Phi_i^n(\mathbf{x}, \mu_i^n, \Sigma_i^n). \tag{8}$$

In the experiments presented in this paper, a white Gaussian noise is used to corrupt the speech signal. It is modeled by only one Gaussian pdf.
In the following Sections, we present three missing feature handling approaches called: estimation by the integrated speech-background model, mean estimation and integration over missing feature bounds.

4.1. Integrated Speech-Background Model

The integrated speech-background model has been successfully used for robust speaker identification [5]. It is based on the fact that the corruption process could be described by a component-wise function of speech and noise. The corrupted speech features are obtained by this function as:

$$y_t = f(s_t, n_t) \tag{9}$$

where s_t and n_t are the feature vectors at frame t of speech and noise, respectively. In the logarithmic domain l, the MAX noise model assumes that the observable feature vector Y^l of noisy speech could be modeled as the maximum of clean speech and background noise vectors, S^l and N^l, respectively [6].

$$Y^l = log(S+N) \approx \max(S^l, N^l) \tag{10}$$

where S and N are the filter-bank energy of speech and noise in the linear domain. In this case, the function $f(\cdot)$ is expressed by the $\max(\cdot)$ operator.
Given the background noise model λ_n, the state i, the clean speaker model λ and the corrupted speech observation, the conditional expectation of clean speech can be expressed as follows:

$$E\{S^l|Y^l, i, \lambda\} = \iint_C S^l p(S^l|Y^l, i, \lambda_n) dS^l dN^l \tag{11}$$

Where C is the contour of integration defined in the two-dimensional space $n-s$ by the $\max()$ operator [5]. If the noise is modeled by one Gaussian pdf, the conditional expectation of the clean speech is given by:

$$\begin{aligned} E\{S^l|Y^l, i, \lambda\} =& p(S^l = Y^l|i, \lambda) Y^l \\ & + (1 - p(S^l = Y^l|i, \lambda)) \\ & \cdot E\{S^l|S^l < Y^l, i, \lambda\} \end{aligned} \tag{12}$$

The integrated background and speech model is described in detail in [5]. By detecting the missing features in noisy speech, we assume that the noise level is known

with high certainty and that the speech components below this level could not be estimated. However, Eq. 12 shows that with the integrated models, the uncertainty about the noise level (noise variance $\neq 0$) is used to derive an estimation of all corrupted speech components including the masked ones. In regions where speech is dominated by noise ($1 - p(S^l = Y^l|i,\lambda) \gg 0$), the expectation of the masked speech components is determined by the term $E\{S^l|S^l < Y^l, i, \lambda\}$. Eq. 13 serves to replace the missing features by the conditional mean of clean speech signal given noisy data

$$E\{S^l|Y^l,\lambda\} = \sum_{i=1}^{M} p_i E\{S^l|Y^l,i,\lambda\}. \qquad (13)$$

A drawback of this method resides in the fact that the speaker models are assumed to have diagonal covariance matrices which is not valid for the filterbank log energies.

4.2. Mean Estimation

In the presence of missing features, the feature vectors $\mathbf{x_t}$ can be composed of two vectors: $\mathbf{x_p}$ and $\mathbf{x_m}$ representing present and missing feature components. The speaker model $\lambda = (\mu_i, \Sigma_i) \quad i = 1, 2, \cdots, M$ can be expressed for each multivariate Gaussian pdf as [7]:

$$\mu = \begin{pmatrix} \mu_p \\ \mu_m \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{pp} & \Sigma_{pm} \\ \Sigma_{mp} & \Sigma_{mm} \end{pmatrix} \qquad (14)$$

The first approach consists of replacing the missing features by the mean of the model or simply by zero. In our case, an additional model with one Gaussian pdf for each speaker is constructed and the feature components with zero values are replaced by their corresponding mean in the model

$$\mathbf{x_m} = \mu_m. \qquad (15)$$

The second approach is based on conditional mean estimation by linear regression. The missing features are estimated as follows:

$$\begin{aligned} \mathbf{x_m} &= E\{\mathbf{x_m}|\mathbf{x_p},\lambda\} \\ &= \mu_m + \Sigma_{mp}(\mathbf{x_p} - \mu_p)\Sigma_{pp}^{-1} \end{aligned} \qquad (16)$$

It should be noted that Eq. 16 requires, for each Gaussian pdf, inverting the matrix Σ_{pp}. Since the positions of present features in the feature vector change from frame to frame, the conditional mean estimation contributes to an excessive computation load. In a large database, the use of this technique becomes too expensive. A common covariance matrix could be used to reduce the computational complexity. In our experiments, an additional multivariate mono-Gaussian model for each speaker with full covariance matrix is computed. The additional model serves for missing features estimation by the calculation of the mean in the first approach and the conditional mean in the second approach.

4.3. Integrating over Missing Feature Bounds

By the marginal distribution use, we assume that the feature components are not bounded and occupy an unlimited space $(-\infty, +\infty)$. However, the feature components are bounded. In a general case, it has been shown that the output probability of GMMs in the framework of Bayesian classification could be expressed as follows [8]:

$$p(\mathbf{x}|\lambda) = p(\mathbf{x_p}|\lambda) \int_{\mathbf{x_l}}^{\mathbf{x_u}} p(\mathbf{x_m}|\mathbf{x_p},\lambda) d\mathbf{x_m} \qquad (17)$$

where $\mathbf{x_l}$ and $\mathbf{x_u}$ are the lower and upper limits of the missing feature space. $p(\mathbf{x_m}|\mathbf{x_p},\lambda)$ is a conditional pdf defined by the conditional mean:

$$\mu_{m|p} = \mu_m + \Sigma_{mp}(\mathbf{x_p} - \mu_p)\Sigma_{pp}^{-1} \qquad (18)$$

and the conditional variance:

$$\Sigma_{m|p} = \Sigma_{mm} - \Sigma_{mp}\Sigma_{pp}^{-1}\Sigma_{pm}. \qquad (19)$$

The possible calculation methods of the integral in Eq. 17 can be found in [8]. A simple method consists of transforming the full covariance matrices to diagonal ones by setting the off diagonal values to zero. The bounds of integration are chosen as follows: as the noise is additif in the spectral domain, the noise masked components have an upper bound equal to the estimated noise energy. The minimal value of speech energy is equal to zero. As the feature components are defined in the log domain, the lower bound of energies is then transformed to $-\infty$. Consequently, Eq. 17 takes the following form:

$$p(\mathbf{x}|\lambda) = p(\mathbf{x_p}|\lambda) \int_{-\infty}^{log(E_m^{noise})} p(\mathbf{x_m}|\lambda) d\mathbf{x_m} \qquad (20)$$

where E_m^{noise} is the estimated noise energy in a masked subband during speech pauses.

5. SIMULATIONS AND RESULTS

From NTIMIT database, 400 speakers are selected for the experiments and for each speaker two sentences are provided for the tests. 32 Gaussian pdfs are chosen to represent the speaker models. The use of marginal distributions is well-suited for partially corrupted speech. However, in order to assess the recognition improvement in the presence of wide-band noise, an artificial white Gaussian noise is added to the speech data at different signal-to-noise ratios (SNRs). The SNR is calculated using only the variance of the speech segments with silence removed. Experiments in "clean" conditions give an equal-error-rate (EER) equal to 12.3% when the models have diagonal covariance matrices. The EER of 9.9% is obtained with full covariance matrices.

As shown in Figure 2, a substantial decrease of the EER is observed with the marginal distributions use comparing to the classical MMSE speech enhancement. The use of the marginal distributions with full covariance matrices is

possible but leads to very high an extreme computation load due to matrice inversions at each frame.

In Figure 3, we compare different missing feature han-

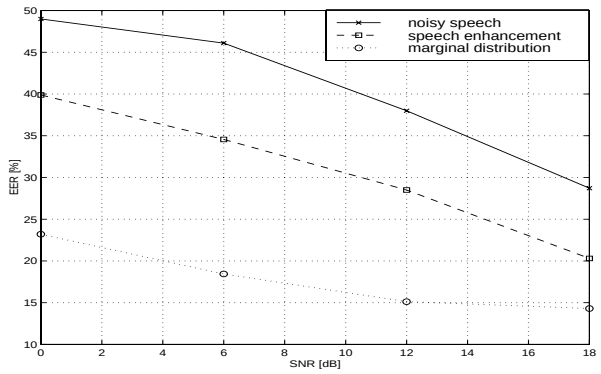

Figure 2: *EER: speaker models are trained with diagonal covariance matrices.*

dling techniques. The integration over missing feature bounds method gives noticeable improvement at low SNR conditions. The estimation by the integrated speech-background model performs better than the mean estimation (unconditional) but less than the marginal distributions and the other techniques.

In Figure 4, results are reported for the mean and the

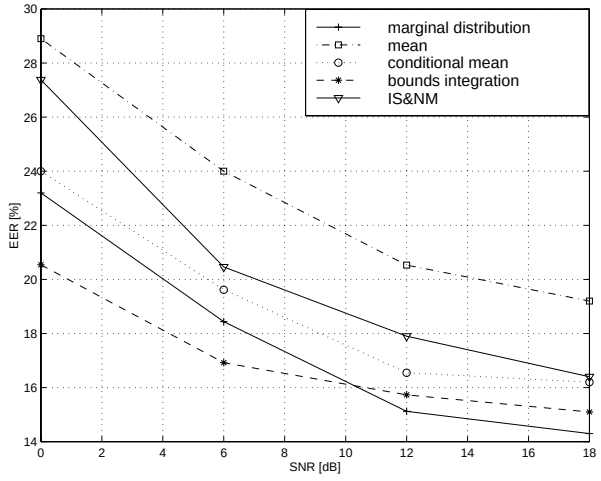

Figure 3: *EER : speaker models are trained with diagonal covariance matrices. IS&NM= integrated speech and noise model for missing features estimation*

conditional mean estimations when the speaker models have full covariance matrices. We observe that such estimations decrease the EER and that the conditional mean estimation gives the best recognition performance. This observation encourages to use orthogonalized feature parameters in order to lower the computation load while using the full covariance matrices.

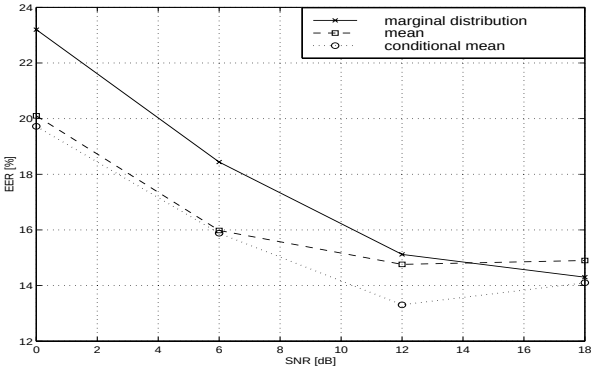

Figure 4: *EER: speaker models are trained with full covariance matrices except for marginal distribution .*

6. CONCLUSIONS

We have presented in this paper an efficient speech enhancement and missing feature detection based on the MMSE spectral estimator. At low SNR conditions and in the presence of wide-band noise, the use of marginal distributions in likelihood calculation led to a significant decrease of the EER. However, including some statistical information about the missing features (e.g. conditional mean estimation) in the recognition task gave an additional increase of the performance.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[2] A. Drygajlo and M. El-Maliki, "Spectral subtraction and missing feature modeling for speaker verification", *in Proc. of Eusipco'98*, pp. 355–358, Rhodes, September 1998.

[3] 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, April 20-23 1998.

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

[5] R.C. Rose, E.M. Hofstetter, and D.A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise", *IEEE Trans. on Speech Audio Processing*, vol. 2, pp. 245–257, April 1994.

[6] A. Nádas, D. Nahamoo, and M.A. Picheney, "Speech recognition using noise-adaptive prototypes", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1495–1502, October 1989.

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

[8] A. Morris, M. Cooke, and P.D. Green, "Some solutions to the missing feature problem in data classificaion, with application to noise robust ASR", *in Proc. of ICASSP'98*, vol. 2, pp. 737–740, Seattle, May 1998.