STATISTICAL ESTIMATION OF UNRELIABLE FEATURES FOR ROBUST SPEECH RECOGNITION

Philippe Renevey and Andrzej Drygajlo

Signal Processing Laboratory,
Swiss Federal Institute of Technology, Lausanne

[Philippe.Renevey,Andrzej.Drygajlo]@epfl.ch

ABSTRACT

This paper addresses the problem of robust speech recognition in noisy conditions in the framework of hidden Markov models (HMMs) and missing feature techniques. It presents a new statistical approach to detection and estimation of unreliable features based on a probabilistic measure and Gaussian mixture model (GMM). In the estimation process, the GMM is compensated using parameters of the statistical model of additive background noise. The GMM means are used to replace the unreliable features. The GMM based technique is less complex than the corresponding HMM based estimation and gives similar improvement in the recognition performance. Once unreliable features are replaced by the estimated clean speech features, the entire set of spectral features can be transformed to the other feature domain characterized by higher baseline recognition rate (*e.g* MFCCs) for final recognition using continuous density hidden Markov models (CDHMMs) with diagonal covariance matrices.

1. INTRODUCTION

Recent works have shown that the application of the missing feature approach in speech and speaker recognition under noisy conditions improves the recognition rates [1–6]. In this approach the time-frequency representation of the noisy speech signal is partitioned into reliable (present) and unreliable (missing) regions according to noise masking criteria. In the framework of this approach two main techniques in the classification process have been implemented: marginalisation - when unreliable data are ignored, data imputation - when unreliable data are estimated.

Our previous works have demonstrated that the missing feature modelling succeeds in speech and speaker recognition when using spectral subtraction techniques not only for speech enhancement but also for missing feature detection purposes [4, 5, 7, 8].

In this paper, we propose a new statistical method for detection and estimation of unreliable features. It is based on a probabilistic measure that signal-to-noise ratio (SNR) is greater than 0 dB and Gaussian mixture model compensation. Noise signals are represented by probability density functions. Features detected as unreliable are enhanced using a statistical spectral subtraction and the unreliable features are replaced by a weighted sum of the GMM means. The advantage of this approach is that the detection and estimation processes can be followed by any automatic speech recognition system with transformed spectral features (*e.g.* cepstral coefficients).

2. DETECTION OF UNRELIABLE FEATURES

Several spectral-subtraction type criteria have been proposed for unreliable feature detection [4,6,9]. In this paper only two of them, negative energy criterion and SNR criterion are presented and used for comparison purposes.

The negative energy criterion

$$|\hat{x}(\omega)|^2 = |x(\omega) + n(\omega)|^2 - |\hat{n}(\omega)|^2 < 0 \tag{1}$$

where $\hat{x}(\omega)$ is the estimate of the clean signal, results from the power spectral subtraction algorithm. In this case, features corresponding to negative energy are declared unreliable.

When using SNR criterion, features are declared unreliable if the estimated SNR is smaller than 0 dB.

It means that

$$\log\left(\frac{|\hat{x}(\omega)|^2}{|\hat{n}(\omega)|^2}\right) < 0 \quad or \quad |\hat{x}(\omega)|^2 < |\hat{n}(\omega)|^2 \tag{2}$$

By adding $|\hat{x}(\omega)|^2$ in both sides of Eq. 2 and using the Cauchy-Schwartz inequality, we obtain

$$|\hat{x}(\omega)|^2 < \frac{1}{2}|x(\omega) + n(\omega)|^2 = \frac{1}{2}|y(\omega)|^2 \tag{3}$$

If we replace $\hat{x}(\omega)$ by $y(\omega) - \hat{n}(\omega)$ under the assumption that $y(\omega) - \hat{n}(\omega) > 0$, the following relation is obtained

$$|y(\omega)| < \frac{\sqrt{2}}{\sqrt{2}-1}|\hat{n}(\omega)| = 3.41|\hat{n}(\omega)| \tag{4}$$

which is equivalent to the criterion based on the general spectral subtraction with subtraction factor $\alpha = 3.41$:

$$|\hat{x}(\omega)| = |x(\omega) + n(\omega)| - \alpha|\hat{n}(\omega)| \tag{5}$$

This solution was presented in [5] and the experimental search for optimal over-subtraction factor gave a value for α close to 3.

In this paper we propose a new type of detector based on statistical distribution of the noise. The noise is considered as being normally distributed and this distributions is estimated during speech pauses. The probability that the SNR is greater than zero is:

$$P(SNR > 0) = \int_{-\infty}^{y(\omega)/2} \frac{1}{\sqrt{2\pi}|\hat{\sigma}_n(\omega)|} e^{\left(\frac{(x-\hat{\mu}_n(\omega))^2}{2\hat{\sigma}_n(\omega)^2}\right)} dx \tag{6}$$

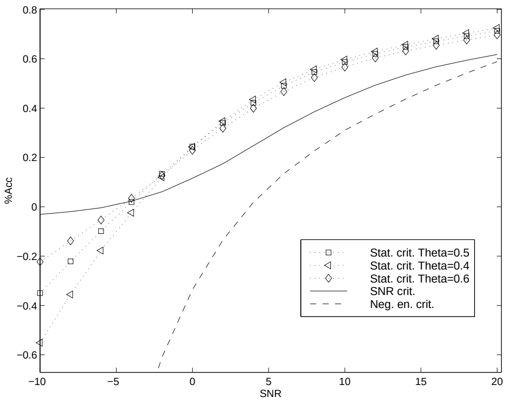

Figure 1: Accuracy for the three detection methods obtained for Lynx helicopter noise from the Noisex'92 database, estimated on 20 digits sequence of the TIDIGITS database.

where $\hat{\mu}_n(\omega)$ and $\hat{\sigma}_n(\omega)^2$ are the estimated mean and variance of the noise for frequency band ω. A feature is considered as unreliable if

$$P(SNR > 0) < \theta \tag{7}$$

where $\theta \in [0, 1]$ is a threshold value.

In order to compare the role of the detection threshold θ, we define a measure for comparing a reference mask (the unreliable data have a SNR lower than 0 dB) with the masks resulting from the negative energy, SNR and probabilistic criteria:

$$\%Corr = \frac{\sum_{t \in T} \sum_{w \in \Omega} R(t, \omega | test) R(t, \omega | ref)}{\sum_{t \in T} \sum_{w \in \Omega} R(t, \omega | ref)} 100\% \tag{8}$$

The percentage of unreliable data labeled as reliable by the detector is subtracted from the percentage of correct detection. An accuracy measure, similar to the one defined for speech recognizer is obtained:

$$\%Acc = \%Corr - \frac{\sum_{t \in T} \sum_{w \in \Omega} R(t, \omega | test) U(t, \omega | ref)}{\sum_{t \in T} \sum_{w \in \Omega} R(t, \omega | ref)} 100\% \tag{9}$$

where T and ω are time and frequency intervals of the time-frequency decomposition of the signal. $R(t, \omega|.) = 1$ if the feature for (t, ω) is declared reliable and zero otherwise. $U(t, \omega|.)$, similarly defined, is equal to one for unreliable features. $test$ and ref refer to the test and the reference detectors.

Fig. 1 presents the accuracy of the three detection methods. We can observe for SNRs greater than -5 dB that the statistical detector outperforms the two other detectors and influence of the threshold value θ on the detection accuracy is not important. The value of θ determines the tradeoff between correctly detected and misclassified reliable and unreliable features.

Fig. 2 shows the differences between the three investigated masks. The detector based on negative energy criterion introduces many misclassified regions. The detector based on the SNR criterion underestimates the reliable features. The probabilistic criterion proposed in this paper yields the closest approximation to the reference mask.

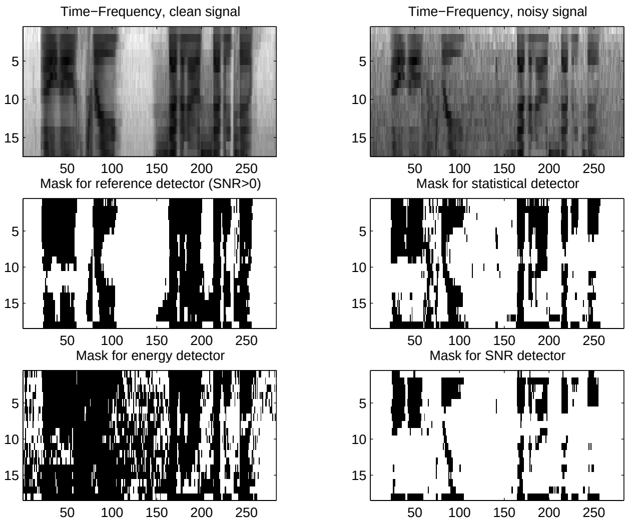

Figure 2: Representation of the masks for the three detectors for factory noise at SNR 10 dB.

3. USING MARGINAL DENSITIES FOR IGNORING UNRELIABLE FEATURES

In HMMs, each state is defined by emission and transition probabilities. For a single state model Γ, the probability to emit vector $X = [x(1), ..., x(\omega), ..., x(\Omega)]^T$ is expressed as

$$Prob\left(X|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega=1}^{\Omega} \Phi\left(x\left(\omega\right), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \tag{10}$$

where p_i is the weight for ith Gaussian pdf, X is a vector containing the log-spectrum components of critical bands and $\mu_{\Gamma_i(\omega)}$, $\sigma^2_{\Gamma_i(\omega)}$ are the mean and variance for ith Gaussian pdf in frequency band ω.

The components of X can be divided into reliable and unreliable features. In Eq.10 the contribution of the reliable and unreliable components can be expressed as follows

$$\begin{aligned} Prob\left(X|\Gamma\right) = \sum_{i=1}^{M} p_i & \prod_{\omega, reliable} \Phi\left(x\left(\omega\right), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \\ & \prod_{\omega, unreliable} \Phi\left(x\left(\omega\right), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \end{aligned} \tag{11}$$

In the previous work [7], only the reliable components were used for the recognition task. In this case, the marginal pdfs are used to compute the emission probabilities.

$$Prob\left(X|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega, reliable} \Phi\left(x\left(\omega\right), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \tag{12}$$

4. ESTIMATION OF UNRELIABLE FEATURES

The data imputation technique gives an estimate of the unreliable parameters. This approach offers the advantage that other than filter bank features can be used, *e.g.* Mel frequency cepstral coefficients (MFCCs) which generally give better baseline recognition results.

A HMMs based data imputation method has been proposed in [10, 11]. The recognized state sequence obtained using only the reliable data is used to impute the unreliable data. The unreliable data are replaced by a weighted sum of the means of the distributions of the recognized state sequence.

When using time-dependent statistical models such as HMMs, the state sequence needed for the estimation of the unreliable data is hidden . If an error in the decoding sequence occurs, it can influence the recognition in the second feature domain. Therefore the method for the estimation of the unreliable data proposed in this paper, is based on time-independent Gaussian mixture models (GMMs) instead of HMMs. It models roughly the distribution of the clean speech data. This is clearly a suboptimal modelization, but sufficient and computationally efficient for data imputation. The GMM likelihood for a vector of parameters $X = [x(1), ..., x(\omega), ..., x(\Omega)]^T$ is defined as:

$$p(X|\lambda) = \sum_{i=1}^{M} P(i) \prod_{\omega=1}^{\Omega} \Phi\left(x(\omega), \mu(\omega,i), \sigma^2(\omega,i)\right) \qquad (13)$$

where M represents the number of Gaussian distributions, $P(i)$, $\mu(\omega,i)$ and $\sigma^2(\omega,i)$ are, respectively, the weighting factor , mean and variance for Gaussian i.

The means and variances of the GMM are compensated to cope with the additive noise, as in parallel model combination (PMC).

The means and the variances of each Gaussian distribution are transformed into the magnitude spectral domain using an inverse log-normal approximation:

$$\mu_{lin}(\omega,i) = \exp\left(\mu(\omega,i) + \frac{\sigma^2(\omega,i)}{2}\right) \qquad (14)$$

$$\sigma^2_{lin}(\omega,i) = \mu^2_{lin}(\omega,i)\left(\exp(\sigma^2(\omega,i)) - 1\right) \qquad (15)$$

These means and variances are modified to include the distortion introduced by the additive noise. Means and variances of the noise are estimated during speech pauses.

$$\tilde{\mu}_{lin}(\omega,i) = \mu_{lin}(\omega,i) + \mu_n(\omega) \qquad (16)$$

$$\tilde{\sigma}^2_{lin}(\omega,i) = \sigma^2_{lin}.(\omega,i) + \sigma^2_n(\omega) \qquad (17)$$

The means and variances of the GMM are transformed back into the log-spectral domain:

$$\tilde{\mu}(\omega,i) = \log\left(\frac{\tilde{\mu}^2_{lin}(\omega,i)}{\sqrt{\tilde{\mu}^2_{lin}(\omega,i) + \tilde{\sigma}^2_{lin}(\omega,i)}}\right) \qquad (18)$$

$$\tilde{\sigma}^2(\omega,i) = \log\left(\frac{\tilde{\sigma}^2_{lin}(\omega,i)}{\tilde{\mu}^2_{lin}(\omega,i)} + 1\right) \qquad (19)$$

Using this noisy GMM, the weighting factor associated with each distribution is computed as follows:

$$\gamma(i|Y,\tilde{\lambda}) = \frac{P(i)\prod_{\omega=1}^{\Omega} \Phi\left(y(\omega), \tilde{\mu}(\omega,i), \tilde{\sigma}^2(\omega,i)\right)}{p(Y|\tilde{\lambda})} \qquad (20)$$

Finally, the reliable data are enhanced using a spectral subtraction and the unreliable features are replaced by a weighted sum of the GMM means:

$$\begin{aligned} \hat{x}_r(\omega) &= \log\left(\exp\left(y_r(\omega)\right) - \mu_n(\omega)\right) \\ \hat{x}_u(\omega) &= \sum_{i=1}^{M} \gamma(i|Y,\tilde{\lambda}).\mu(\omega,i) \end{aligned} \qquad (21)$$

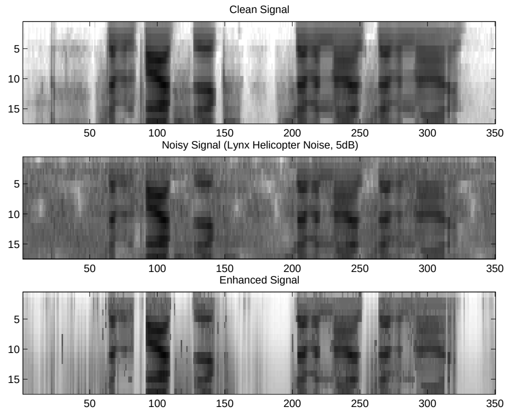

Figure 3: Clean signal, noisy signal and enhanced signal

Fig. 3 presents the features spectra obtained as a result of the speech enhancement based on the proposed method for Lynx helicopter noise from the Noisex'92 database (digit sequence "75679-79").

5. EXPERIMENTS AND RESULTS

The recognition system was developed using Hidden Markov Toolkit (HTK). The time-frequency representation of the signal is provided by the 17 bands Bark filter bank analysis. Digits models have been trained on TIDIGITS database down-sampled to a frequency of 8kHz. TIDIGITS database was used for the digits recognition experiments using 224 utterances from 152 speakers extracted randomly from the database. Noises from the NOISEX database have been added to obtain test utterances.

Recognition experiments have been performed using the Bark filter bank features and Mel frequency cepstral coefficients (MFCCs) with energy and first and second derivatives. The MFCCs are calculated from the filter bank features.

The HMM based recognition system with three detectors based on negative energy, SNR and probabilistic criteria was tested using:

- marginal densities for ignoring unreliable features (Table 1),
- estimation of unreliable features in the filter bank domain (Table 2),
- estimation of unreliable features in the filter bank domain and their transformation into Mel frequency cepstral coefficients (MFCC) domain (Table 3).

The results obtained for data imputation were compared with baseline recognition performance and recognition results when using the general spectral subtraction pre-processing.

All the results show that the recognition with statistical detection and estimation of unreliable features outperforms the two other techniques for all three experiments. The results are given for the optimized values of θ.

Marginal Densities

Method→	Neg. En.	SNR	Stat. Det.	Stat. Det.
			($\theta = 0.7$)	($\theta = 0.2$)
SNR↓	% Acc.	% Acc.	% Acc.	% Acc.
-5	-7.21	13.46	**16.07**	13.07
0	14.73	22.70	**28.44**	22.64
5	40.94	38.07	40.88	**44.90**
10	62.37	47.07	53.89	**58.93**
15	69.26	58.55	64.48	**70.85**
20	78.32	69.77	72.32	**78.06**
25	80.36	74.23	73.47	**83.61**

Table 1: Recognition results for Lynx helicopter noise using only the reliable features.

Bark Filter Bank Parameters

Method→	Base	GSS	Neg. En.	SNR	Stat. Det.
					($\theta = 0.7$)
SNR↓	% Acc.	% Acc.	% Acc.	% Acc.	% Acc.
-5	8.67	**13.97**	4.08	6.70	7.65
0	6.38	24.68	21.68	24.49	**27.23**
5	14.67	37.95	42.86	43.24	**45.34**
10	34.18	48.87	59.50	62.12	**64.16**
15	52.17	62.44	68.75	72.00	**75.89**
20	68.24	71.94	74.11	78.89	**81.89**
25	78.12	80.55	78.12	82.82	**85.50**

Table 2: Recognition results for Lynx helicopter noise using estimation of unreliable features in the filter bank domain.

MFCC + E + Δ + $\Delta\Delta$

Method→	Base	GSS	Neg. En.	SNR	Stat. Det.
					($\theta = 0.7$)
SNR↓	% Acc.	% Acc.	% Acc.	% Acc.	% Acc.
-5	-19.32	-41.33	9.63	**26.79**	24.74
0	8.16	2.42	11.16	45.28	**48.66**
5	45.47	51.66	20.47	68.49	**73.66**
10	76.21	82.78	34.18	86.35	**88.27**
15	88.52	90.69	48.60	90.75	**92.35**
20	93.69	92.92	60.40	94.02	**95.41**
25	**96.68**	95.73	71.43	95.73	96.56

Table 3: Recognition results for Lynx helicopter noise in the MFCC domain using estimated unreliable features from the filter bank domain.

6. CONCLUSION

In this paper, a new statistical method for detection and GMM based estimation of unreliable features in noisy speech is proposed. The assessment results show a significant increase in performance of the HMM based recognition system when using a probabilistic criterion for unreliable feature detection in comparison with implementation of the negative energy and SNR criteria. The data imputation technique presented in this paper also opens the possibility of extension of the missing feature techniques to recognition systems with other than spectral domain features, for example MFCCs.

7. REFERENCES

[1] Cooke, M., Green, P., and Crawford, M., “Handling missing data in speech recognition”, *in Int. Conf. on Spoken Language Processing (ICSLP)*, pp. 1555–1558, 1994.

[2] Lippmann, R. P. and Carlson, B. A., “Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise”, *in Eurospeech*, vol. 1, pp. 37–40, Rhodes, Greece, Sep. 1997.

[3] Cooke, M., Morris, A., and Green, P., “Missing data techniques for robust speech recognition”, *in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 863–866, 1997.

[4] Drygajlo, A. and El-Maliki, M., “Speaker verification in noisy environments with combined spectral subtraction and missing feature theory”, *in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–124, 1998.

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

[6] Vizinho, A., Green, P., Cooke, M., and Josifovski, L., “Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: an integrated study”, *in Eurospeech*, vol. 5, pp. 2407–2410, 1999.

[7] El-Maliki, M., Renevey, P., and Drygajlo, A., “Rehaussement par soustraction spectrale et compensation des paramètres manquants pour la reconnaissance robuste du locuteur et de la parole”, *in JEP’98 (Journée d’Étude de la Parole)*, pp. 409–412, Martigny, 1998.

[8] Renevey, P. and Drygajlo, A., “Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition”, *in Eurospeech*, vol. 6, pp. 2627–2630, 1999.

[9] Drygajlo, A. and El-Maliki, M., “Spectral subtraction and missing feature modeling for speaker verification”, *in Proc. of EUSIPCO’98*, pp. 355–358, Rhodes, September 1998.

[10] Morris, A. C., Cooke, M. P., and Green, P. D., “Some solution to the missing feature problem in data classification, with application to noise robust ASR”, *in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 737–740, 1998.

[11] Josifovski, L., Cooke, M., Green, P., and Vizinho, A., “State based imputation of missing data for robust speech recognition and speech enhancement”, *in Eurospeech*, vol. 6, pp. 2833–2836, 1999.