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

MISSING FEATURE THEORY AND PROBABILISTIC ESTIMATION OF CLEAN SPEECH COMPONENTS FOR ROBUST SPEECH RECOGNITION

Philippe Renevey and Andrzej Drygajlo

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

10.21437/Eurospeech.1999-579

ABSTRACT

In the framework of Hidden Markov Models (HMMs), this paper presents a new approach towards robust speech recognition in adverse conditions. The approach is based on statistical modeling of noise by Gaussian distributions and an estimation of idealized clean speech directly in the probabilistic domain using a statistical spectral subtraction method and missing feature compensation. The missing feature approach in the probabilistic domain allows the speech features masked by noise to be dynamically detected and estimated in probability calculations performed in HMM based recognizers. The method combines the advantages of two techniques: the first based on the statistical compensation similar to the parallel model combination and the second one issued from the missing feature theory.

1. INTRODUCTION

Human auditory system has the ability to extract the informative features from speech signal in adverse environment and to take decisions about the reliability of a particular part of the signal. This explains its remarkable robustness to noise. In automatic speech recognition, some efforts for extracting noise-free features and determining the reliability of the data are ongoing [1–7]. The parameters used in speech recognition can be divided in two subsets, the *reliable* or *present* parameters that contain the information about the clean speech signal, and the *unreliable*, masked by noise or *missing* parameters.

The method presented in this paper combines the statistical estimation of the clean speech parameters using statistical noise models and the missing feature theory. The estimate of any noise parameter is not considered as a plain value, but as a probability density function (pdf). In this paper, the estimated noise parameters are considered as normally distributed in each frequency band.

Such an approach is justified by the fact that, even in the case of stationary noise, the measured short-term noise parameters can vary around their mean values from frame to frame. This approach also allows the statistical variation of the noise to be compensated in statistical noisy speech models.

The normal distributions of the noise parameters permit to estimate the normal distributions of clean speech parameters (informative pdfs) or detect the unreliable (missing) data using a novel technique of statistical spectral subtraction. When a parameter is declared missing or masked by noise, it is considered as being uniformly distributed within a determined interval (non informative pdf).

Finally, in the HMMs the emission probabilities are obtained by the integration of the product of estimated idealised clean speech parameters pdfs and the model pdfs.

2. ESTIMATION OF CLEAN SPEECH PARAMETERS AS PROBABILITY DENSITY FUNCTIONS

When speech signal is disturbed by an additive noise, this noise is generally considered as being also additive in the spectral magnitude domain. This is an approximation because the phase difference between the noise and the speech signal is, generally, not equal to zero. Nevertheless, this assumption is useful for speech enhancement algorithms [8–10]. Under such an assumption of additivity, the clean speech magnitude can be expressed as the difference between the noisy speech magnitude and the estimated value of the noise magnitude:

$$|\hat{X}(\omega)| = |Y(\omega)| - |\hat{N}(\omega)| \qquad (1)$$

where $\hat{X}(\omega)$ is the estimated clean speech in the frequency band ω, $Y(\omega)$ is the noisy speech and $\hat{N}(\omega)$ is the noise estimate. Eq.(1) represents the classical spectral subtraction [9]. If $|\hat{N}(\omega)|$ is represented as a probability density function (pdf), $|\hat{X}(\omega)|$ becomes a pdf too.

The distribution of $|\hat{N}(\omega)|$ is considered as normal

$$f\left(|\hat{N}(\omega)|, n\right) = \Phi\left(\mu_{|\hat{N}(\omega)|}, \sigma^2_{|\hat{N}(\omega)|}\right) \qquad (2)$$

where $f(A, \ )$ represents the pdf for parameter A and $\Phi\left(\mu_A, \sigma^2_A\right) = \frac{1}{\sqrt{2\pi}|\sigma_A|} \exp\left(\frac{(a-\mu_A)^2}{2\sigma^2_A}\right)$ is a normal distribution.

From Eqs (1) and (2) the distribution of the estimate of the clean speech parameter is

$$f\left(|\hat{X}(\omega)|, x\right) = \Phi\left(\mu_{|\hat{X}(\omega)|}, \sigma^2_{|\hat{X}(\omega)|}\right) \qquad (3)$$

with $\mu_{|\hat{X}(\omega)|} = |Y(\omega)| - \mu_{|\hat{N}(\omega)|}$ and $\sigma^2_{|\hat{X}(\omega)|} = \sigma^2_{|\hat{N}(\omega)|}$. Generally, speech recognizers achieve better performances in the log-spectral domain than in the spectral domain [11]. The pdf of the estimated clean speech parameters in the log domain is

$$f\left(\hat{X}^l(\omega), x\right) = \frac{\exp(x)}{\sqrt{2\pi}|\sigma_{|\hat{X}(\omega)|}|} \exp\left(\frac{\left(\exp(x) - \mu_{|\hat{X}(\omega)|}\right)^2}{2\sigma^2_{|\hat{X}(\omega)|}}\right) \quad (4)$$

where $\hat{X}^l(\omega) = log\left(|\hat{X}(\omega)|\right)$.
This equation is relatively complex and computationally inefficient. A normal approximation is often used for log-normal distributions [12]. Using such an approximation, Eq.(4) becomes

$$f\left(\hat{X}^l(\omega), x\right) \cong \Phi\left(x, \mu_{\hat{X}^l(\omega)}, \sigma^2_{\hat{X}^l(\omega)}\right)$$
$$\mu_{\hat{X}^l(\omega)} = \log\left(\frac{\mu^2_{|\hat{X}(\omega)|}}{\sqrt{\mu^2_{|\hat{X}(\omega)|} + \sigma^2_{|\hat{X}(\omega)|}}}\right) \quad (5)$$
$$\sigma^2_{\hat{X}^l(\omega)} = \log\left(\frac{\sigma^2_{|\hat{X}(\omega)|}}{\mu^2_{|\hat{X}(\omega)|}} + 1\right)$$

Eq.(5) gives the normal distribution of the estimated clean parameter in the log-spectral domain.

3. MISSING COMPONENTS

When noise totally masks the clean speech parameter, it becomes impossible to estimate the normal distributions of the clean speech parameters. In this case, the masked (missing) parameters are considered as being uniformly distributed between zero and the noisy value.

$$f\left(\hat{X}(\omega), x\right) = \begin{cases} \frac{1}{|Y(\omega)|} & \text{if } 0 \le x \le |Y(\omega)|; \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In the log-spectral domain, Eq.(6) becomes

$$f\left(\hat{X}^l(\omega), x\right) = \begin{cases} \frac{\exp(x)}{|Y(\omega)|} & \text{if } -\infty \le x \le Y^l(\omega); \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Under the hypothesis of additive noise in the magnitude domain, the clean speech parameters cannot be greater than the noisy one. The parameters are declared missing when the mean of the estimated distribution of the noise parameter is greater than the noisy parameter $\mu_{|\hat{N}(\omega)|} > |Y(\omega)|$.

4. PROBABILITY COMPUTATION

In HMMs, each state is defined by emission and transition probabilities. For a singular state model Γ, the probability to emit vector $\hat{\mathbf{X}}^l$ is expressed as

$$Prob\left(\hat{\mathbf{X}}^l|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega=1}^{\Omega} \Phi\left(\hat{X}^l(\omega), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \quad (8)$$

where p_i is the weight for ith Gaussian pdf, $\hat{\mathbf{X}}^l = \left[\hat{X}^l(1) \ldots \hat{X}^l(\omega) \ldots \hat{X}^l(\Omega)\right]^T$ a vector containing the log-spectrum components of critical bands and $\mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}$ the mean and variance for ith Gaussian pdf in frequency band ω.
The components of $\hat{\mathbf{X}}^l$ can be divided into present and missing features. In Eq.(8) the contribution of the present and missing components can be expressed as follows

$$Prob\left(\hat{\mathbf{X}}^l|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega,present} \Phi\left(\hat{X}^l(\omega), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \prod_{\omega,missing} \Phi\left(\hat{X}^l(\omega), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \quad (9)$$

In the previous work [5], only the reliable components (present) are used for the recognition task. In this case, the marginal pdfs are used to compute the emission probabilities.

$$Prob\left(\hat{\mathbf{X}}^l|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega,present} \Phi\left(\hat{X}^l(\omega), \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \quad (10)$$

In Eqs.(9) and (10), the values of $\hat{\mathbf{X}}^l$ are considered as deterministic. In the method proposed in this paper, the clean parameters are represented as pdfs. Consequently Eq.(8) becomes

$$Prob\left(f\left(\hat{\mathbf{X}}^l, \mathbf{x}\right)|\Gamma\right) = \sum_{i=1}^{M} p_i \prod_{\omega=1}^{\Omega} \int_{a(\omega)}^{b(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) f\left(\hat{X}^l(\omega), x\right) dx \quad (11)$$

where $f\left(\hat{\mathbf{X}}^l, \mathbf{x}\right)$ is a vector whose components are the distributions of the estimated clean speech parameters $f\left(\hat{X}^l(\omega), x\right)$, in each frequency band ω.
The present and missing parameters are represented by normal and uniform distributions, respectively. In the case of present parameters (normal distribution), the integral in

Eq.(11) becomes

$$\begin{aligned}
&\int_{a(\omega)}^{b(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) f\left(\hat{X}^l(\omega), x\right) dx \\
&= \int_{a(\omega)}^{b(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \Phi\left(x, \mu_{\hat{X}^l(\omega)}, \sigma^2_{\hat{X}^l(\omega)}\right) \\
&= \frac{\exp\left(-\frac{\left(\mu_{\Gamma_i(\omega)} - \mu_{\hat{X}^l(\omega)}\right)^2}{2\left(\sigma^2_{\Gamma_i(\omega)} + \sigma^2_{\hat{X}^l(\omega)}\right)}\right)}{\sqrt{2\pi\left(\sigma^2_{\Gamma_i(\omega)} + \sigma^2_{\hat{X}^l(\omega)}\right)}} \\
&\quad \int_{a(\omega)}^{b(\omega)} \Phi\left(x, \mu_{\hat{X}^l\Gamma_i(\omega)}, \sigma^2_{\hat{X}^l\Gamma_i(\omega)}\right) dx
\end{aligned} \tag{12}$$

The integral in Eq.(11) represents the integral of a Gaussian distribution with the mean and variance as in Eqs (13) and (14) respectively, multiplied by a constant term.

$$\mu_{\hat{X}^l\Gamma_i(\omega)} = \frac{\mu_{\hat{X}^l(\omega)}\sigma^2_{\Gamma_i(\omega)} + \mu_{\Gamma_i(\omega)}\sigma^2_{\hat{X}^l(\omega)}}{\sigma^2_{\Gamma_i(\omega)} + \sigma^2_{\hat{X}^l(\omega)}} \tag{13}$$

$$\sigma^2_{\hat{X}^l\Gamma_i(\omega)} = \frac{\sigma^2_{\Gamma_i(\omega)}\sigma^2_{\hat{X}^l(\omega)}}{\sigma^2_{\Gamma_i(\omega)} + \sigma^2_{\hat{X}^l(\omega)}} \tag{14}$$

If the interval of integration is from $a(\omega) = -\infty$ to $b = (\omega)\,\infty$, the integral of Eq.(12) becomes a constant term. In the case of missing parameters (uniform distributions), the integral in Eq.(11) becomes:

$$\begin{aligned}
&\int_{a(\omega)}^{b(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) f\left(\hat{X}^l(\omega), x\right) dx \\
&= \int_{-\infty}^{Y^l(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) \frac{\exp(x)}{|Y(\omega)|} dx \\
&= \frac{\exp\left(\mu_{\Gamma_i(\omega)} + \frac{\sigma^2_{\Gamma_i(\omega)}}{2}\right)}{|Y(\omega)|} \\
&\quad \int_{-\infty}^{Y^l(\omega)} \Phi\left(x, \mu_{\Gamma_i(\omega)} + \sigma^2_{\Gamma_i(\omega)}, \sigma^2_{\Gamma_i(\omega)}\right) dx
\end{aligned} \tag{15}$$

Eqs (12) and (15) allow the probability of emission of a vector containing both uniform and normal pdfs to be calculated.

5. EXPERIMENTS

The recognition system was developed using Hidden Markov Toolkit (HTK). The selection of the present/missing parameters is done for each frame of the speech data. The noise is estimated during speech pauses. The present parameters are the components whose magnitude exceeds the mean of the estimated noise. The recognizer uses Bark filter bank analysis to obtain feature parameters. Seventeen frequency bands have been used.

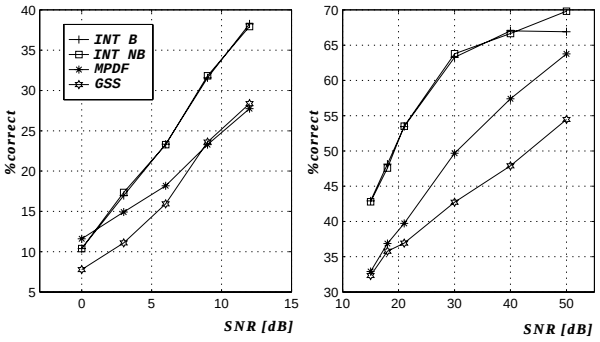

Figure 1: Recognition results for speech-like noise

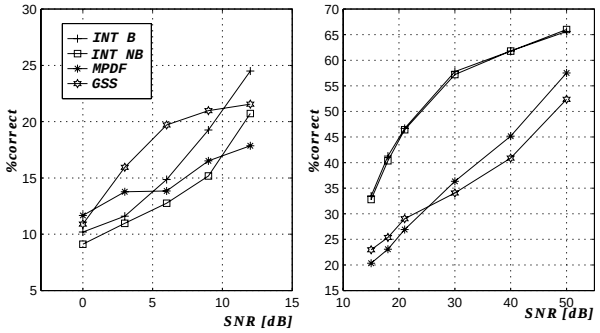

Figure 2: Recognition results for white Gaussian noise

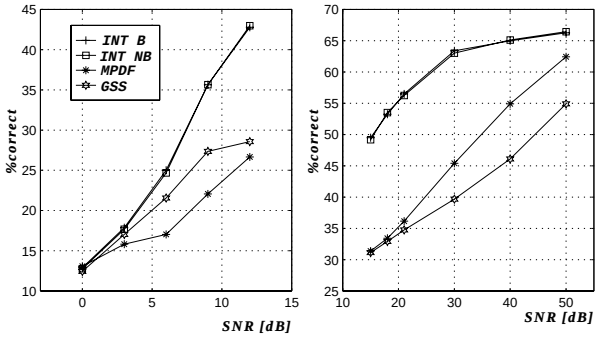

Figure 3: Recognition results for Lynx helicopter noise

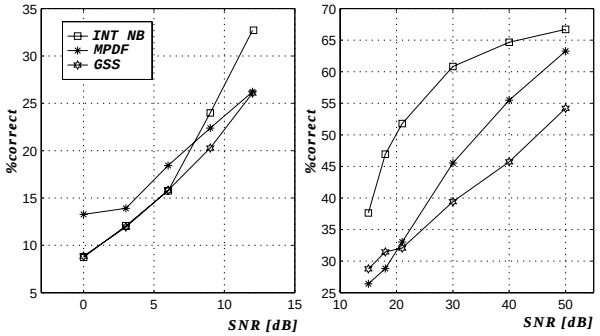

Figure 4: Recognition results for factory noise

Fifty five phoneme models have been trained on TIMIT database down-sampled to a frequency of 8kHz. TIDIGIT database has been 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. The speech intensity is measured as root mean square (RMS) of all segments with an magnitude exceeding 10 % of the maximum magnitude of the signal. Therefore speech pauses (of the clean signal) do not alter the speech intensity measurement.

The recognition tests have been performed for four kinds of compensation scheme using:

1. **GSS:** General spectral subtraction with $\alpha = 1.5$ and $\beta = 0.002$ [8, 9, 13].
2. **MPDF:** Missing feature compensation using marginal probability density function (Eq.(10)).
3. **INTB:** Integration of the pdfs over bounded interval. The values of $a(\omega)$ and $b(\omega)$ are respectively $-\infty$ and $Y^l(\omega)$.
4. **INTNB:** Integration of the pdfs over infinite interval. The values of $a(\omega)$ and $b(\omega)$ are respectively $-\infty$ and ∞.

Four kinds of noises have been used for tests:

- **Speech like noise:** A noise with a spectral shape similar to that of speech signals (Fig. 1).
- **White Gaussian noise:** (Fig. 2).
- **Lynx noise:** Noise in the cockpit of a Lynx helicopter (Fig. 3).
- **Factory noise:** Mechanical noise in a factory (Fig. 4).

The recognition results, presented in Figs.1-4, show that the proposed method achieves better performance than the other tested methods. The interval of integration has negligible influence on the recognition performances (differences between **INTB** and **INTNB**). The most significant improvement of the recognition rate is obtained for SNRs from 9 to 40 dB.

6. CONCLUSIONS

In this paper, a new method combining missing feature theory and statistical estimation of the clean speech parameters have been proposed. A significant improvement in the speech recognition performance has been achieved when compared with other methods such as general spectral subtraction and missing feature compensation using marginal pdfs. The interval of integration of Eq.(12) has a negligible influence on the recognition results while the infinite interval integration reduces the computational load.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] Berouti, M., Schwartz, R., and Makhoul, J., "Enhancement of speech corrupted by acoustic noise", *in Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 208–211, 1979.

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

[10] Gales, M. F. J., *Model-based techniques for noise robust speech recognition*, PhD thesis, University of Cambridge, 1994.

[11] Erell, A. and Weintraub, M., "Filterbank energy estimation for recognition of noisy speech", *IEEE Trans. Speech Audio Process.*, vol. 1, pp. 68–76, 1993.

[12] Crow, E. L. and Shimizu, K., *Lognormal distributions: theory and applications*, Marcel Dekker, 1988.

[13] Lockwood, P. and Boudy, J., "Experiments with a non linear spectral subtraction (NSS) and hidden Markov models and projection for robust speech recogntion in cars", *Speech Communication*, vol. 11, pp. 215–228, 1992.