FEATURE COMPENSATION USING MORE ACCURATE STATISTICS OF MODELING ERROR

Woohyung Lim, Jong Kyu Kim and Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Seoul 151-742, Korea
E-mail: {whlim, ckkim}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

In this paper, we propose a novel approach to feature compensation for robust speech recognition in noisy environments. We analyze the statistics of the modeling error in the log Mel magnitude spectrum domain, and model it as a Gaussian distribution. The mean and variance of the distribution are Gaussian functions of the SNR, which enables us to use the SNR dependency of the modeling error efficiently. The proposed feature compensation approach, which is based on the interacting multiple model (IMM) technique, incorporates the statistics of the modeling error and shows significant improvement in the AURORA2 speech recognition task.

Index Terms— Feature compensation, robust speech recognition, modeling error statistics

1. INTRODUCTION

The performance of speech recognition systems drops dramatically in the presence of background noise. One of the effective way to cope with this performance degradation is the feature compensation technique in which noisy input speech features are compensated before being decoded by the recognition models trained on clean speech. Two main approaches to feature compensation include stereo training data based approach [1] and speech corruption model based approach [2]-[6]. In the former approach, the relation of the clean speech, noise, and distorted speech is modeled by end-to-end results, and the complex process of speech distortion does not matter in explicit manner. But there is a weakness to get the appropriate stereo data, and it is not easy in the practical point of view. On the other hand, in the latter approach, how the speech is distorted is modeled by explicit expression, so the accurate model of the speech distortion should be established for the performance improvement. But because there are some nuisance parameters that we cannot achieve from the given noisy speech, such as the phase relationship between the clean speech and the noise, it is difficult to treat the accurate model.

One of the successful techniques for solving shortcomings of these approaches is to use the phase information explicitly. In [2], a deterministic approach was proposed, which is related to the nonlinear spectral subtraction. On the other hand in [3] and [4], a statistical model of the phase related modeling error was proposed in the name of the ignorance model [3] or the phase-sensitive model [4]. In [3] and [4], the relationship among the log spectra of the clean speech, noise, and noisy speech was described by not a conventional deterministic function but a statistical model. The characteristics of the modeling error was analyzed in the log Mel power spectrum domain, and applied to the MMSE estimator to achieve great improvement. But to avoid the implementation complexity, these algorithms ignored the dependency between the modeling error and the SNR, which is needed for modeling the speech corruption more accurately. Especially when we use the log Mel magnitude spectrum, which is made by taking logarithm to the Mel filter banks of the magnitude spectrum instead of the power spectrum, the relationship among the clean speech, noise, and noisy speech features is more complicated including the phase information, and the SNR dependency becomes more important because the mean of the modeling error cannot be assumed zero any more.

In this paper, we propose a new feature compensation technique, called the interacting multiple model (IMM) with modeling error statistics (IMM-MES) approach, in which the error of the speech corruption model is statistically incorporated into the original IMM algorithm [5],[6]. The IMM-MES algorithm makes an assumption that the modeling error can be treated as a random variable with a normal distribution whose mean and variance are well-approximated by Gaussian functions of the signal-to-noise ratio (SNR). From a number of speech recognition experiments on AURORA2 database under the condition of clean training, we have been able to find that the proposed approach improves the performance of the original IMM technique.

2. MODELING ERROR STATISTICS

In this section, let us consider how to derive the speech corruption model in the feature vector domain. Let $X[k]$, $N[k]$, and $Z[k]$ denote the kth discrete Fourier transform (DFT) coefficients of the clean speech, background noise and noisy speech, respectively. Then, their relation is described as

$$Z[k] = X[k] + N[k]. \tag{1}$$

It is noted that (1) is equivalent to the relation in the original waveform domain and is an exact model for speech corruption.

Our purpose is to express (1) in the feature vector domain. Here, we will derive the speech corruption model in terms of the log Mel-scale filter bank outputs which are widely applied for speech recognition. If $\tilde{Z}$ represents the output of a Mel-scale filter, it can be written by

$$\begin{aligned}\tilde{Z} &= \sum_k W_k |Z[k]| \\ &= \sum_k W_k \left(|X[k]|^2 + |N[k]|^2 + 2|X[k]||N[k]| \cos\theta_k \right)^{1/2} \end{aligned} \tag{2}$$

where θ_k denotes the angel between the two complex numbers $X[k]$ and $N[k]$, and W_k is the non-negative gain for the kth spectral component. As shown in (2), each Mel-scale filter output is given as a

linear combination of the spectral magnitudes. One may define the Mel-scale filter output in a different way such as the weighted combination of the squared spectral magnitudes. We can also define the Mel-scale filter outputs corresponding to the clean speech and noise in a similar way as follows:

$$\tilde{X} = \sum_k W_k |X[k]| \ , \ \tilde{N} = \sum_k W_k |N[k]|. \tag{3}$$

It is important to show the relationship among these filter bank outputs. But (2) is too complicated to describe it, we use the following indirect approach.

Since $-1 \leq \cos\theta_k \leq 1$, $\tilde{Z}$ is bounded such that

$$\begin{aligned}
&\left| \sum_k W_k |X[k]| - \sum_k W_k |N[k]| \right| \\
&\quad \leq \sum_k W_k \big| |X[k]| - |N[k]| \big| \\
&\quad \leq \tilde{Z} \\
&\quad \leq \sum_k W_k \big| |X[k]| + |N[k]| \big| \\
&\quad = \left| \sum_k W_k |X[k]| + \sum_k W_k |N[k]| \right|.
\end{aligned} \tag{4}$$

Using (3), we can rewrite (4) as

$$\left| \tilde{X} - \tilde{N} \right| \leq \tilde{Z} \leq \left| \tilde{X} + \tilde{N} \right|. \tag{5}$$

If we introduce a phase related variable Θ ($-1 \leq \Theta \leq 1$), (5) can be described in terms of a parametric function given by

$$\tilde{Z} = \left(\tilde{X}^2 + \tilde{N}^2 + 2\tilde{X}\tilde{N}\Theta \right)^{1/2}. \tag{6}$$

Based on (6), the log spectral component of the noisy speech z is obtained as follows:

$$\begin{aligned}
z &= \log \tilde{Z} \\
&= \log[\tilde{X} + \tilde{N}] + \log \left[\frac{\tilde{X}^2 + \tilde{N}^2 + 2\tilde{X}\tilde{N}\Theta}{\left(\tilde{X} + \tilde{N} \right)^2} \right]^{1/2} \\
&= x + \log[1 + \exp(n - x)] \\
&\quad + \log \left[1 + \Omega \frac{\exp(x - n)}{\left(1 + \exp(x - n) \right)^2} \right]^{1/2}
\end{aligned} \tag{7}$$

where

$$\Omega = 2(\Theta - 1) \tag{8}$$

and x and n are the corresponding log spectral components of the clean speech and noise, respectively. The right hand side of (7) can be separated into two parts such that

$$z = z_c + z_p \tag{9}$$

with

$$\begin{aligned}
z_c &= x + \log[1 + \exp(n - x)] \\
z_p &= \log \left[1 + \Omega \frac{\exp(x - n)}{\left(1 + \exp(x - n) \right)^2} \right]^{1/2}.
\end{aligned} \tag{10}$$

If we are given the log spectra of the noise and clean speech, z_c is completely determined. In contrast, z_p depends not only on x and n but also on the phase related parameter Ω which is not directly available in the log spectral domain. Since our purpose is to build a speech corruption model in the log spectral domain, z_c can be considered an approximated model with z_p being treated as the modeling error.

The difficulty of modeling lies on the fact that the modeling error z_p can not be obtained solely from x and n. In order to cope with this problem, we employ a statistical model for z_p. Let

$$\epsilon(\eta, \Omega) = -z_p = -\log \left[1 + \Omega \frac{\exp(\eta)}{\left(1 + \exp(\eta) \right)^2} \right]^{1/2} \tag{11}$$

where $\eta = x - n$. Then, we can find that $\epsilon(\cdot, \cdot)$ is a function of both the SNR η and the phase related variable Ω. For a given value of η, we assume that $\epsilon(\cdot, \cdot)$ is a random variable with a Gaussian distribution such that

$$p(\epsilon|\eta) = \mathcal{N}(\epsilon; \mu_\epsilon, \Sigma_\epsilon) \tag{12}$$

in which μ_ϵ and Σ_ϵ are the mean and variance, respectively. Because the modeling error $\epsilon(\cdot, \cdot)$ depends on the SNR, the parameters μ_ϵ and Σ_ϵ should also be given as functions of η. Since $\exp(\eta)/(1 + \exp(\eta))^2$ in (11) is a bell shape function which has the maximum value at $\eta = 0$, μ_ϵ and Σ_ϵ can be well approximated by Gaussian functions with respect to η. From a number of simulations on artificial speech corruption with a variety of noise sources, we can find that the Gaussian functions represent the modeling error distribution well. Some of the results are in Fig. 1.

Now, we can describe μ_ϵ and Σ_ϵ as follows:

$$\begin{aligned}
\mu_\epsilon(\eta) &= E[\epsilon(\eta, \Omega)|\eta] = \alpha \exp\left(-\frac{\eta^2}{2\beta} \right) \\
\Sigma_\epsilon(\eta) &= \text{var}[\epsilon(\eta, \Omega)|\eta] = \gamma \exp\left(-\frac{\eta^2}{2\delta} \right)
\end{aligned} \tag{13}$$

where $\{\alpha, \beta, \gamma, \delta\}$ are positive valued parameters that can be estimated from a set of training data. If η is not a constant but a random variable distributed according to a Gaussian probability density function (pdf) with mean μ_η and variance Σ_η, we can further extend (13) as follows:

$$\begin{aligned}
\tilde{\mu}_\epsilon &= E[\epsilon(\eta, \Omega)|\mu_\eta, \Sigma_\eta] \\
&= \alpha \left(\frac{\beta}{\beta + \Sigma_\eta^2} \right) \exp\left(-\frac{\mu_\eta^2}{2(\beta + \Sigma_\eta)} \right) \\
\tilde{\Sigma}_\epsilon &= \text{var}[\epsilon(\eta, \Omega)|\mu_\eta, \Sigma_\eta] \\
&= \gamma \left(\frac{\delta}{\delta + \Sigma_\eta} \right)^{1/2} \exp\left(-\frac{\mu_\eta^2}{2(\delta + \Sigma_\eta)} \right) \\
&\quad + \alpha^2 \left(\frac{\beta}{\beta + 2\Sigma_\eta} \right)^{1/2} \exp\left(-\frac{\mu_\eta^2}{\beta + 2\Sigma_\eta} \right) \\
&\quad - \alpha^2 \left(\frac{\beta}{\beta + \Sigma_\eta} \right) \exp\left(-\frac{\mu_\eta^2}{\beta + \Sigma_\eta} \right).
\end{aligned} \tag{14}$$

Because not only the deterministic value of the noise but also the statistics of it is estimated in the IMM algorithm, (14) will play an important role to improve the feature compensation performance.

3. IMM WITH MODELING ERROR STATISTICS

In this section, we propose the IMM-MES technique to compensate the noisy feature vectors. The proposed approach modifies the original IMM algorithm [5],[6] with a more sophisticated speech corruption model. In the IMM-MES technique, the pdf of the clean speech

is assumed to be a Gaussian mixture distribution given by

$$p(\mathbf{x}) = \sum_{k=1}^{M} p(k)\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{15}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_D]'$ is a log spectral vector of the clean speech with the prime denoting the vector transpose. In (15), M is the total number of mixture components and $p(k)$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ represent the given weight, mean and covariance of the kth Gaussian, respectively.

In order to make the nonlinear relationship among z, x and n, which is shown in (7), a tractable one, the deterministic part z_c for each Mel-filter output is approximated by a piecewise linear model given by

$$\begin{aligned} z_c &= x + \log[1 + \exp(n - x)] \\ &\approx A_k x + B_k n + C_k \end{aligned} \tag{16}$$

if x is assumed to have come from the kth mixture component. The coefficients $\{A_k, B_k, C_k\}$ are obtained by the statistical linear approximation (SLA) algorithm [7]. The evolving environment model in conjunction with the piecewise linear observation model described in (16) enables us to form a state space model for each mixture component as follows:

$$\begin{aligned} n(t) &= n(t-1) + w(t) \\ z(t) &= A_k x(t) + B_k n(t) + C_k - \epsilon(\eta(t), \Omega(t)) \end{aligned} \tag{17}$$

where t represents a specific time index and $w(t)$ is a zero-mean white Gaussian process. In (17), the noise feature $n(t)$ is treated as the state variable, and it is assumed to be distributed according to a Gaussian pdf $\mathcal{N}(n(t); \mu_n(t), \Sigma_n(t))$ with $\mu_n(t)$ and $\Sigma_n(t)$ denoting the mean and covariance at time t. For a detailed information of the IMM algorithm, interested readers are referred to [5],[6].

The parameters $\{\mu_n(t), \Sigma_n(t)\}$ concerned with the background noise are sequentially estimated by the IMM algorithm that consists of four steps summarized in the following [5],[6]:

- *Mixing step*: the state estimates obtained from each cluster in the previous time are combined together to produce a single set of state estimates, which is provided to each Kalman filter as an initial statistic.
- *Kalman step*: the conventional Kalman update is carried out conditioned on the initial estimates computed from the *Mixing step*.
- *Probability computation step*: the a posteriori probability associated with each cluster is updated.
- *Output generation step*: the state estimates are generated by combining the estimates of all the clusters. In our case, this step is the same to the *Mixing step*.

If we use the statistics of the modeling error, the *Kalman* and the *probability computation* steps of the original IMM algorithm should be modified. In the *Kalman step*, the innovation of the kth mixture component at time t, $e(t|k)$, is now computed as

$$\begin{aligned} e(t|k) &= z(t) - \mu_z^p(t|k) \\ &= z(t) - A_k\mu_k - B_k\mu_n^p(t|k) - C_k + E[\epsilon(\eta(t), \Omega(t))] \end{aligned} \tag{18}$$

and further, its covariance as

$$\begin{aligned} R_e(t|k) = {} & B_k\Sigma_n^p(t|k)B_k' + A_k\Sigma_k A_k' \\ & + \mathrm{var}[\epsilon(\eta(t), \Omega(t))] + 2\mathrm{cov}[z_c, \epsilon(\eta(t), \Omega(t))] \end{aligned} \tag{19}$$

where $\mu_n^p(t|k)$ and $\Sigma_k^p(t|k)$ are respectively the mean and covariance of the one-step-ahead predictive state estimate in the kth mixture component at time t, and $\mathrm{cov}[z_c, \epsilon(\eta(t), \Omega(t))]$ represents a cross-covariance between z_c and $\epsilon(\cdot, \cdot)$. If we assume that $x(t)$ and $n(t)$ are mutually uncorrelated random processes with distributions given by

$$\begin{aligned} x(t) &\sim \mathcal{N}(x(t); \mu_k, \Sigma_k) \\ n(t) &\sim \mathcal{N}(n(t); \mu_n^p(t|k), \Sigma_n^p(t|k)), \end{aligned} \tag{20}$$

$E[\epsilon(\eta(t), \Omega(t))]$ and $\mathrm{var}[\epsilon(\eta(t), \Omega(t))]$ can be derived from (14) with $\mu_\eta = \mu_k - \mu_n^p(t|k)$ and $\Sigma_\eta = \Sigma_k + \Sigma_n^p(t|k)$. In addition, $\mathrm{cov}[z_c, \epsilon(\eta(t), \Omega(t))]$ is ignored in this work due to its smaller absolute value compared to other components. Based on (18) and (19), the Kalman gain $K_f(t|k)$ is obtained as follows:

$$K_f(t|k) = \Sigma_n^p(t|k)B_k' R_e^{-1}(t|k). \tag{21}$$

After the *Kalman step*, we conduct the *probability calculation step* in which the posterior probability corresponding to each mixture component is updated. Since the mixture component is assumed to be independent of the previous observations, we have

$$p(k|\mathbf{Z}_t) = \frac{p(\mathbf{z}(t)|k, \mathbf{Z}_{t-1})p(k)}{p(\mathbf{z}(t)|\mathbf{Z}_{t-1})} \tag{22}$$

where $\mathbf{z}(t)$ is a Dth order log spectral vector of the noisy speech at time t and $\mathbf{Z}_t$ denotes the feature vector sequence $\{\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(t)\}$. After estimating the parameters concerned with the background noise, the clean speech estimate is computed according to the minimum-mean-square error (MMSE) criterion.

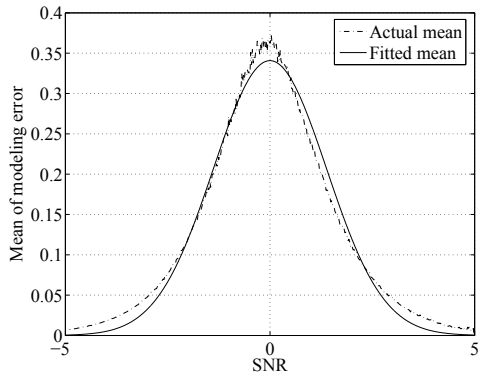
4. EXPERIMENTAL RESULTS

Performance of the IMM-MES algorithm was evaluated on the AURORA2 database which consists of the TI-DIGITS data down-sampled to 8 kHz [8]. The AURORA2 database is regarded as the clean speech data and it has been artificially contaminated by adding the noises recorded under several conditions. Three sets of speech database were prepared for the recognition experiments. In test set A, the four noises (subway, babble, car and exhibition hall) were added to the clean data at SNR's of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. In test set B, another four different noises (restaurant, street, airport and train station) were added to the clean data at the same SNR's. Finally in test set C, two of the noises from set A and set B (subway and street) were added to the clean data and there also existed a channel mismatch. Results were presented as an average performance in five SNR conditions from 20dB to 0dB.

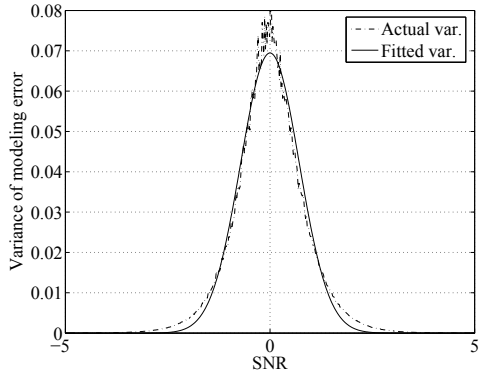
Feature compensation was performed in the log spectral domain, and the compensated log spectra were converted to the cepstral coefficients through discrete cosine transform (DCT). Before applying the IMM-MES algorithm, the input signal was passed through the signal bias removal (SBR) algorithm [9] for removing the possible channel mismatch. In the IMM-MES algorithm, clean speech log spectra were modeled by a mixture of 128 Gaussian distributions with diagonal covariance matrices. In order to estimate the relevant parameters for the statistics of modeling error $\{\alpha, \beta, \gamma, \delta\}$ for each Mel-filter output, we added randomly generated noises to the clean speech waveforms and fitted the Gaussian curves to the sample mean and variance using the steepest-descent algorithm [10]. Some of the estimated parameters are shown in Table 1 and the fitted curves for a log Mel-filter output are plotted in Fig. 1 in conjunction with the corresponding empirical statistics. In Fig. 1, we can see that the

Table 1. The estimated parameters for modeling error statistics of the 5th and the 19th Mel-filter outputs.

	α	β	γ	δ
Low freq. (5th)	0.34	1.92	0.069	0.51
High freq. (19th)	0.32	2.04	0.024	0.75



(a) Mean



(b) Variance

Fig. 1. Mean and variance values obtained with the fitted curve for the 5th Mel-filter output.

mean and variance of the modeling error become larger as the SNR approaches 0 dB. From the simulation results, we can conclude that the Gaussian curve provides a close approximation.

The recognition results obtained from the AURORA2 task in clean training condition are shown in Table 2 where the relative improvement represents an averaged word recognition error reduction rate compared to the baseline over the SNR range from 20 dB to 0 dB. From the results, we can easily observe that the IMM-MES approach outperformed the conventional IMM algorithm which does not consider the modeling error information.

5. CONCLUSIONS

In this paper, we have analyzed the statistics of the modeling error in the log Mel magnitude spectrum. Based on the analysis, we have proposed a new feature compensation technique incorporated with the statistical model of the modeling error. Proposed statistical model represents the modeling error characteristics of the SNR dependency and we have found that the modeling error where the noise and the speech level are similar is not negligible. From a number of experiments, our approach have been shown to improve the recognition performance efficiently.

Table 2. Word accuracies(%) over AURORA2 database for clean training condition (Relative improvements(%) compared to the baseline system).

	set A	set B	set C	Average
Baseline	61.34	55.75	66.14	60.06
IMM	81.09	81.59	77.57	80.59 (48.82)
IMM-MES	83.28	84.36	79.26	82.91 (55.41)

6. ACKNOWLEDGEMENTS

This work was partly supported by IT R&D Project funded by Korean Ministry of Information and Communication.

7. REFERENCES

[1] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. D. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. ICASSP*, vol. 1, pp. 301-304. May 2001.

[2] Q. Zhu and A. Alwan, "The effect of additive noise on speech amplitude spectra: A quantitative analysis," *IEEE Signal Processing Letters*, vol. 9, pp. 275-277, Sep. 2002.

[3] L. Deng, J. Droppo, and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," in *Proc. ICASSP*, vol. 1, pp. 829-832, May 2002.

[4] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech, Audio Process.*, vol. 12, no. 2, pp. 133-143, Mar. 2004.

[5] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Processing Letters*, vol. 5, no. 6, pp. 146-149, June 1998.

[6] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Commun.*, vol. 37, no. 4, pp. 231-248, July 2002.

[7] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Processing Letters*, vol. 5, no. 1, pp. 8-10, Jan. 1998.

[8] H. -G. Hirsch and D. Pearch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, pp. 16-20, Oct. 2000.

[9] M. G. Rahim and B. -H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech, Audio Process.*, vol. 4, no. 1, pp. 19-30, Jan. 1996.

[10] S. Haykin, *Adaptive Filter Theory*, 4th ed. New Jersey: Prentice-Hall, 2002.