Clean Speech Feature Estimation based on Soft Spectral Masking

Young Joon Kim[1,2], *Woohyung Lim*[2],*and Nam Soo Kim*[2]

[1]Electronics and Telecommunications Research Institute, Deajeon, Korea
[2]School of Electrical Engineering and INMC, Seoul National University, Seoul, Korea
kjun@etri.re.kr, whlim@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

In this paper, we first analyze the problems of speech and noise contamination process in noise-masking point of view, and propose a new approach to estimate degree of noise masking effect on clean speech distribution model based on sequential noise estimation. Sequential noise estimation is performed frame-by-frame using interacting multiple model (IMM) algorithm, so that real-time implementation is possible. After applying IMM algorithm, degree of noise masking effect named as noise masking probability(NMP) is calculated. Estimation of clean speech spectrum in noisy environments is performed by controlling the advantages of log spectrum domain and those of linear spectrum domain algorithm based on NMP. We have performed recognition experiments under noise conditions using the AURORA2 database which is developed for a standard reference of speech recognition performance. Simulation results show that this approach is effective when noise masking effect is dominated at low SNR.

Index Terms: speech recognition, feature compensation, noise masking probability.

1. Introduction

Noise degrades significantly the performance of speech recognition system running in real conditions. The performance of speech recognition systems degrades seriously if there exist background noise, channel distortion, acoustic echo or a variety of interfering signals. So one of the key issues in practical speech recognition is to achieve robustness against the mismatch between the training and testing environments [1]. Methods for obtaining noise robustness are classified into two types. One adapts the acoustic models in the recognizer to any kinds of noises based on model adaptation techniques [2]. The major disadvantage for these kinds of methods is that they need a huge computational load with large vocabulary speech recognition systems. The other is enhancing the feature vectors based on noise reduction techniques before they are fed into the recognizer [3]. An easiest way to alleviate the recognition performance degradation is to employ a feature compensation technique in which the input speech features are compensated before being decoded by the recognition models trained on clean speech. In this paper, we will focus on latter approach to obtain robustness against the environment in which the clean speech is corrupted by any background noise.

In the context of feature compensation, a variety of approaches have been developed. One of the successful approaches to feature compensation applies piecewise linear approximation to the speech contamination procedure defined in the feature vector domain [4, 5]. Even though this approach has been found effective in high signal-to-noise ratio (SNR) regions, it usually causes much erroneous estimates when the instantaneous SNR is low.[6]. In order to overcome this problem, Kim et al. proposed an approach which combines two separate compensation techniques [7]: the spectral subtraction (SS) and interacting multiple model (IMM) algorithms. In this approach, which is referred to as the soft decision IMM (SDIMM) algorithm, the speech absence probability (SAP) is computed by the SS module, and it is applied to control the switching between the two compensation methods. The IMM algorithm utilizing a detailed clean speech distribution dominates the estimation in the active speech regions while the SS algorithm provides a stationary estimate during the non-speech periods.

Since, however, SAP is computed by the SS method which does not have any prior knowledge of the clean speech distribution, it is likely to make a wrong decision especially when either the noise or speech characteristics are time-varying. For that reason, it is considered desirable to compute the SAP by means of an algorithm which employs a more precise model for clean speech feature distribution. Moreover, it is generally known that the feature vector components which are masked by the noise are responsible for the performance degradation of the speech recognizer in adverse environments. In this paper, we analyze the problems of estimation from noise-masking clusters and propose a novel approach to measure the noise masking effect by taking advantage of a detailed clean speech distribution. For each feature vector component, we compute the noise masking probability (NMP) which accounts for how much the current component is masked by the noise. Once NMP's are computed, our approach to feature compensation is operating in a similar way to that of the SDIMM technique. For each feature vector component, the two clean speech estimates obtained by both the Wiener filter and the IMM algorithm are linearly interpolated with the NMP being treated as the interpolating weight.

2. Feature Compensation based on IMM

Let $\mathbf{z} = [z_1, z_2, \cdots, z_D]^t$ represent a D-dimensional log spectrum of the noisy input with ^{t} denoting the transpose. Then,

$$z_d = x_d + \log\left[1 + \exp\left(n_d - x_d\right)\right] \ , \quad \text{for } d = 1, 2, \cdots, D \qquad (1)$$

where $\mathbf{x} = [x_1, x_2, \cdots, x_D]^t$ and $\mathbf{n} = [n_1, n_2, \cdots, n_D]^t$ are the log spectra of the clean speech and added noise, respectively. In the IMM technique, the distribution of $\mathbf{x}$ is described in terms of a Gaussian mixture model (GMM) as follows:

$$p(\mathbf{x}) = \sum_{k=1}^{M} p(k)\mathcal{N}\left(\mathbf{x}; \mu_k, \Sigma_k\right) \qquad (2)$$

where M is the total number of mixture components, and $p(k)$, μ_k and Σ_k represent the *a priori* probability, mean vector and covariance matrix of the kth Gaussian distribution, respectively. On the other hand, the log spectrum of the background noise is assumed to be distributed according to a single Gaussian given by

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}). \tag{3}$$

To make the nonlinear function given in (1) mathematically tractable, the IMM algorithm applies a linear approximation such that

$$\mathbf{z} = A_k\mathbf{x} + B_k\mathbf{n} + C_k \tag{4}$$

if $\mathbf{x}$ is assumed to have come from the kth mixture component. Here, the coefficient matrices $\{A_k, B_k, C_k\}$ are obtained by the statistical linear approximation (SLA) technique which is based on Taylor series expansion of a nonlinear function [5]. In order to estimate the evolving characteristics of the noise, current noise is assumed to vary slowly from the previous time and described by

$$\mathbf{n}_{t+1} = \mathbf{n}_t + \mathbf{w}_t \tag{5}$$

where $\mathbf{w}_t$ represents a Gaussian process of zero mean and covariance Q. Equation (4) and (5) construct a linear state space model and we can apply the IMM algorithm to estimate sequentially environmental parameter, $\lambda_{\mathbf{n}} = \{\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}}\}$. IMM technique consists of four major steps [4].

- Step 1) *Mixing step* : the parameter estimates of the noise obtained from each mixture component are combined together to produce a single noise estimate.

$$\begin{aligned}
\mu^0_{\mathbf{n}}(t-1|k) &= E[\mathbf{n}_{t-1}|k_t = k, Z_{t-1}] \\
&= \sum_{j=1}^{M} \gamma_k(t-1)\, \hat{\mu}_{\mathbf{n}}(t-1|j) \\
\Sigma^0_{\mathbf{n}}(t-1|k) &= Cov[\mathbf{n}_{t-1}|k_t = k, Z_{t-1}] \\
&= \sum_{j=1}^{M} \gamma_k(t-1)\, [\hat{\Sigma}_{\mathbf{n}}(t-1|j) + \\
&\quad (\hat{\mu}_{\mathbf{n}}(t-1|j) - \hat{\mu}^0_{\mathbf{n}}(t-1|j)) \\
&\quad (\hat{\mu}_{\mathbf{n}}(t-1|j) - \hat{\mu}^0_{\mathbf{n}}(t-1|j))^T]
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mu}_{\mathbf{n}}(t-1|j) &= E[\mathbf{n}_{t-1}|k_{t-1} = j, \mathbf{Z}_{t-1}] \\
\hat{\Sigma}_{\mathbf{n}}(t-1|j) &= Cov[\mathbf{n}_{t-1}|k_{t-1} = j, \mathbf{Z}_{t-1}] \\
\gamma_j(t-1) &= p(k_{t-1} = j|\mathbf{Z}_{t-1}).
\end{aligned}$$

- Step 2) *Kalman step* : the conventional Kalman update is carried out conditioned on the initial estimates computed from the *Mixing Step*.
  - one-step-ahead predictive state estimate (time update)

$$\begin{aligned}
\mu^p_{\mathbf{n}}(t|j) &= \hat{\mu}^0_{\mathbf{n}}(t-1) \\
\Sigma^p_{\mathbf{n}}(t|j) &= \hat{\Sigma}^0_{\mathbf{n}}(t-1) + Q.
\end{aligned}$$

  - Innovation and its covariance

$$\begin{aligned}
e(t|j) &= \mathbf{z}_t - A_j\mu_j - B_j\mu^p_{\mathbf{n}}(t|j) - C_j \\
R_e(t|j) &= B_j\Sigma^p_{\mathbf{n}}(t|j)B_j^T + A_j\Sigma_jA_j^T.
\end{aligned}$$

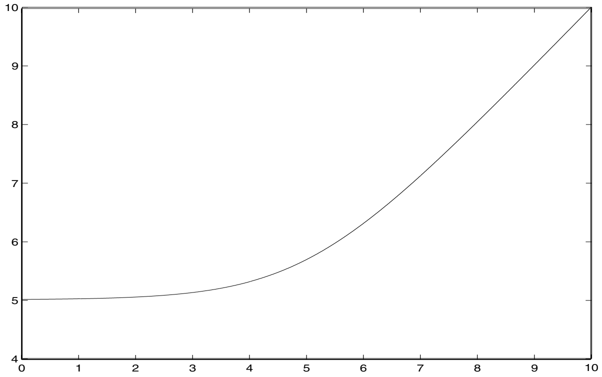

Figure 1: Plot of function $z = x + \log\left[1 + \exp\left(n - x\right)\right]$. n= 5.0 and x ranges from 0 to 10

  - Kalman Gain

$$\begin{aligned}
K_f(t|j) &= \Sigma^p_{\mathbf{n}}(t|j)B_j^TR_e^{-1}(t|j) \\
K^*_f(t|j) &= \alpha K_f(t|j).
\end{aligned}$$

  - Correction (measurement update)

$$\begin{aligned}
\hat{\mu}_{\mathbf{n}}(t|j) &= \mu^p_{\mathbf{n}}(t|j) + K^*_f(t|j)e(t|j) \\
\hat{\Sigma}_{\mathbf{n}}(t|j) &= \Sigma^p_{\mathbf{n}}(t|j) - K^*_f(t|j)B_j\Sigma^p_{\mathbf{n}}(t|j).
\end{aligned}$$

- Step 3) *Probability calculation step* : the posteriori probability associated with each mixture component is updated.

$$\gamma_j(t) = \frac{p(\mathbf{z}_t|k_t = j, \mathbf{Z}_{t-1})\; p(k_t = j)}{p(\mathbf{z}_t|\mathbf{Z}_{t-1})}.$$

- Step 4) *Output generation step* : the noise parameter estimates are generated by combining the estimates of all the mixture components.

$$\begin{aligned}
\hat{\mu}_{\mathbf{n}}(t) &= \sum_{j=1}^{M} \gamma_j(t)\, \hat{\mu}_{\mathbf{n}}(t|j) \\
\hat{\Sigma}_{\mathbf{n}}(t) &= \sum_{j=1}^{M} \gamma_j(t)\, [\hat{\Sigma}_{\mathbf{n}}(t|j) + (\hat{\mu}_{\mathbf{n}}(t|j) - \hat{\mu}_{\mathbf{n}}(t)) \\
&\quad (\hat{\mu}_{\mathbf{n}}(t|j) - \hat{\mu}_{\mathbf{n}}(t))^T]
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mu}_{\mathbf{n}}(t) &= E[\mathbf{n}_t|k_t = j, \mathbf{Z}_t] \\
\hat{\Sigma}_{\mathbf{n}}(t) &= Cov[\mathbf{n}_t|k_t = j, \mathbf{Z}_t] \\
\gamma_j(t) &= p(k_t = j|\mathbf{Z}_t).
\end{aligned}$$

3. Noise Masking Effect on Clean Speech Model

In order to illustrate the contamination relationship of speech and noise in log spectrum domain, we plot (1) in Figure 1 as a function of x_d keeping n_d fixed. It can be seen from this figure that when speech component x_d is much smaller than noise component n_d, the function (1) outputs n_d only. These functional relationship explain the noise masking effects of n_d on x_d. When noise masking effects dominate, the exact estimation of n_d cannot lead to exact estimation of x_d because speech information is masked by noise.

In techniques based on *a priori* clean speech model such as GMM, a procedure of linearization is unavoidable in order to have a computationally tractable model. Without considering masking effects, however, the composed noisy distribution using linear approximation is quite different from the true distribution of noisy speech. This difference causes a dramatic bias on the estimation of clean speech even when the true distribution of clean speech and noise is known. Clean speech estimation from clusters, most of which are masked by noise, would rather deteriorate the estimation performance than improve it.

To cope with this problem, we first discriminate 'noise masking cluster' in which speech are masked by noise. Based on the parameters of the GMM and the estimated noise statistics, discrimination between noise masking one and the other can be made using the fact that dth dimension of kth cluster noisy component is not sensitive to the speech cluster variation. This means that the derivative of a function (1) with respect to speech component represents little change in that function.

Let $\mu_k = [\mu_{k,1}, \mu_{k,2}, \cdots, \mu_{k,D}]^t$ denote the mean vector of the kth Gaussian of the GMM associated with the clean speech and $\hat{\mu}_{\mathbf{n}} = [\hat{\mu}_{\mathbf{n},1}, \hat{\mu}_{\mathbf{n},2}, \cdots, \hat{\mu}_{\mathbf{n},D}]^t$ be the estimate for $\mu_{\mathbf{n}}$ obtained from the IMM algorithm. Then, we can classify all the mixture components of the GMM into two disjoint subsets. Let $M_{m,d}$ be the set of indices corresponding to the mixture components where the dth spectral element is found masked by the noise, and $M_{o,d}$ denote its complementary set. Then, the dth element of the kth mixture component is decided to be masked by the noise if

$$\left.\frac{\partial z_d}{\partial x_d}\right|_{x_d=\mu_{k,d}, n_d=\hat{\mu}_{\mathbf{n},d}} = \frac{1}{1+\exp(\hat{\mu}_{\mathbf{n},d} - \mu_{k,d})} < \eta \qquad (6)$$

where η is a small positive threshold and z_d is defined in (1).

4. Clean Feature Estimation based on Spectral Masking

In this section, we introduce NMP which accounts for how unreliable the estimate from the IMM algorithm would be. NMP is defined separately for each element of the feature vector. Let us denote the NMP associated with the dth element of the input noisy feature vector $\mathbf{z}$ by $NMP_d(\mathbf{z})$. Then,

$$\begin{aligned} NMP_d(\mathbf{z}) &= p(k \in M_{m,d}|\mathbf{z}) \\ &= \frac{p(\mathbf{z}|k \in M_{m,d})}{p(\mathbf{z})} \\ &= \frac{p(\mathbf{z}|k \in M_{m,d})}{p(\mathbf{z}|k \in M_{m,d}) + p(\mathbf{z}|k \in M_{o,d})} \qquad (7) \end{aligned}$$

where $\hat{\lambda}_{\mathbf{n}}$ is the estimate for $\lambda_{\mathbf{n}}$ obtained from the IMM algorithm, and $p(k)$ and $p(\mathbf{z}|k, \hat{\lambda}_{\mathbf{n}})$ represent the *a priori* probability and the likelihood, respectively, of the kth mixture component.

Based on method to identify the mixture components which are masked by the noise proposed in previous section, we can estimate the contributions of noise masking clusters and those of its complementary set by

$$p(\mathbf{z}|k \in M_{m,d}) = \sum_{k \in M_{m,d}} p(k)p(\mathbf{z}|k, \hat{\lambda}_{\mathbf{n}}) \qquad (8)$$

$$p(\mathbf{z}|k \in M_{o,d}) = \sum_{k \in M_{o,d}} p(k)p(\mathbf{z}|k, \hat{\lambda}_{\mathbf{n}}) \qquad (9)$$

Once the NMP's for all the feature vector elements are computed, we can estimate the clean speech log spectrum in a way similar to the SDIMM technique [7]. Let $\hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_D]^t$ denote the estimate for the clean speech feature vector $\mathbf{x}$ when the input is $\mathbf{z}$. Then,

$$\hat{x}_d = NMP_d(\mathbf{z})\hat{x}_d^{wiener} + (1 - NMP_d(\mathbf{z}))\,\hat{x}_d^{imm} \qquad (10)$$

in which $\hat{x}_d^{wiener}$ and $\hat{x}_d^{imm}$ represent the two separate estimates for x_d provided by the Wiener filter and IMM algorithms, respectively. In this paper, to obtain $\hat{x}_d^{wiener}$, we apply the first-stage mel-warped Wiener filter algorithm proposed in [8] without the voice activity detector (VAD). The frequency response of the Wiener filter is smoothed and time-warped using a filter bank that incorporates 23 mel-scale bins.

5. Experimental results

The proposed algorithm was evaluated on the AURORA2 task in which the database consists of the TI-DIGITS data downsampled to 8 kHz [11] [9]. 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 have been prepared for the recognition experiments. In test set A, the four noises (suburban train, babble, car and exhibition hall) are added to the clean data at SNR's of 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. In test set B, another four different noises (restaurant, street, airport and train station) are added to the clean data at the same SNR's. Finally in test set C, two of the noises of set A (subway and street) are added and there also exists a channel mismatch. Results are presented as an average value for five SNR conditions from 20dB to 0dB.

The baseline recognition system was built based on a set of continuous density Gaussian mixture hidden Markov models (HMM's). There were eleven digit models with sixteen states, one silence model with three states and one short pause model with one state. Training and testing were performed using the HTK software [11]. Speech features for recognition consisted of twelve cepstral coefficients derived from 23 mel-spaced triangular filter outputs and log energy, and these thirteen parameters were augmented with the corresponding delta and acceleration coefficients. All the HMM's were trained in the clean training condition. 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).

	set A	set B	set C	Average
Baseline	61.34	55.75	66.14	60.06
IMM only	80.69	81.35	76.23	80.06
IMM+SAP	80.94	81.96	77.15	80.59
IMM+NMP	83.77	83.90	78.80	82.83

Table 1: *Word accuracies (%) in clean training condition.*

In our experiment, the decision threshold η defined in (6) was fixed to 0.1. A typical example of NMP obtained from a noisy input with the corresponding clean speech log spectral energy is shown in Fig. 2 where we also plot the trajectory of the SAP computed by the SS module. It can be shown from this figure that

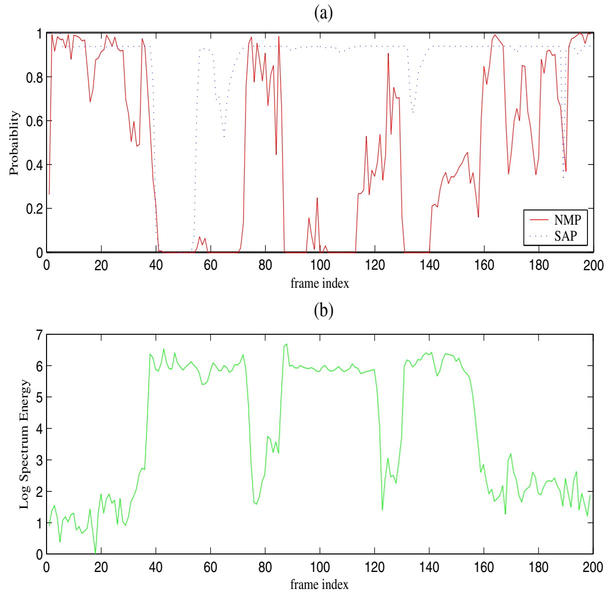

Figure 2: Comparison of NMP and SAP in car noise condition at 10dB SNR. (a) NMP and SAP estimated from a noisy speech signal (b) Corresponding clean speech log spectral energy

NMP fits more closely to the clean speech energy variation than the SAP. For the purpose of comparison, we also tried an approach where (7) is replaced by

$$\hat{x}_d = SAP_d(\mathbf{z})\hat{x}_d^{wiener} + (1 - SAP_d(\mathbf{z}))\hat{x}_d^{imm} \tag{11}$$

with $SAP(\mathbf{z})$ denoting the local SAP provided by the SS technique. The recognition results obtained from the AURORA2 task in clean training condition are shown in Table 1 where the word accuracies averaged over the SNR range from 0-20 dB are listed. In Table 1, IMM+NMP and IMM+SAP represent the proposed algorithms which are based on NMP and SAP, respectively. From the result, it is apparent that IMM+NMP produced higher recognition accuracy than both IMM and IMM+SAP over all the SNR ranges. IMM+NMP improved the performance of the IMM algorithm up to 14.71 %. Figure 3 show experimental results for test data contaminated by babble and car noise condition. As shown in this Figure, the proposed algorithm is more effective in low SNR.

6. Conclusions

In this paper, we have proposed a new approach to measure the relative degree of speech activity. Proposed NMP has been applied to feature compensation and found to improve the overall recognition performance particularly in low SNR conditions. Proposed NMP is very useful method applicable to 'missing feature theory' and 'feature domain compensation' based on clean speech model.

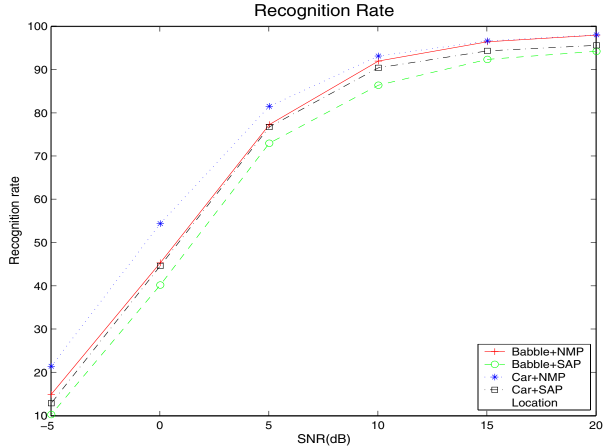

Figure 3: Recognition Result in Babble and Car Noise condition

7. References

[1] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*. MA: Kluwer Academic Press, 1996.

[2] M. J. F. Gales and S. J. Young, "Robust continuous speech recogntion using Parallel Model Combination," *IEEE Trans. on SAP.*, vol.4, no.5, pp. 352-359, 1996.

[3] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 217-220, Sept. 2001.

[4] N. S. Kim, "Feature domain compensation of nonstationary noise for robust speech recognition," *Speech Communication*, vol.37, pp.231-248, July 2002.

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

[6] J. C. Segura, M. C. Benitez, A. D. Torre, S. Dupont and A.J.Rubio, "VTS residual noise compensation," *International Conference on Acoustics, Speech, and Signal Processing*, pp.409-412, May. 2002.

[7] N. S. Kim, Y. J. Kim, and H. W. Kim, "Feature compensation based on soft decision," *IEEE Signal Processing Letters*, vol.11, no.3, pp.378-381, Mar. 2004.

[8] A. Agarwal and Y. M. Cheng, "Two-stage Mel-Warped Wiener filter for robust speech recognition," Proc. IEEE ASRU workshop, 1999.

[9] http://www.icp.inpg.fr/ELRA/home.html.

[10] H. -G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. Int. Conf. Spoken Language Processing*, pp.16-20, October 2000.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK book - version3.0*. July 2000.