Feature Compensation Technique for Robust Speech Recognition in Noisy Environments

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

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

Abstract

In this paper, we analyze the problems of the existing interacting multiple model (IMM) and spectral subtraction (SS) approaches and propose a new approach to overcome the problems of these algorithms. Our approach combines the IMM and SS techniques based on a soft decision for speech presence. Results reported on AURORA2 database show that proposed approach shows 14.26 % of average relative improvement compared to the IMM algorithm in the speech recognition experiments.

1. Introduction

One of the key issues in practical speech recognition is to achieve robustness against the mismatch between the training and testing environments [1]. The performance of speech recognition systems degrades seriously if there exist background noise, channel distortion, acoustic echo or a variety of interfering signals. In this paper, we will focus on the environment in which the clean speech is corrupted by any background noise.

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 general, the feature compensation techniques are classified into two categories where one is the *task-independent* family and the other is the *task-dependent* family. A *task-dependent* technique requires a specific clean speech distribution model which is usually trained based on a set of clean speech data [2], [3]. Since the vocabularies are usually limited in a speech recognition task, the data used for training the clean speech model are collected within the domain of recognition. Even though the *task-dependent* approach shows a dramatic improvement due to the sophisticated modelling of clean speech distribution, the performance is likely to degrade for the out-of-vocabulary speech. For that reason, in order to take advantage of the strength of the *task-dependent* approach, the clean speech distribution model needs to be re-trained whenever the recognition task changes. On the other hand in the *task-independent* approach, we do not need to train the clean speech distribution model, and it can be easily applied to any speech recognizer as a pre-processing unit [4]-[6]. A simplified model for clean speech distribution is adopted in the *task-independent* approach and the relevant parameters are estimated on-line during the feature compensation procedure.

One of the successful *task-dependent* techniques is the interacting multiple model (IMM) approach [3], which is found to significantly reduce the recognition error in slowly-evolving noise environments. The IMM approach approximates the speech contamination process in terms of a piecewise linear model, and estimates the time-varying noise characteristic sequentially by means of a bank of multiple Kalman filters. Examining a large amount of recognition error patterns, we have discovered that the IMM approach causes many insertion errors while it reduces the substitution and deletion errors quite a lot. This phenomenon confirms that the non-speech periods do not fit well to the approximated model provided by IMM.

Spectral subtraction (SS) is practically the predominant speech enhancement technique and it can be applied as a *task-independent* feature compensation module [4]-[6]. In the SS technique, a gain function is defined in terms of the clean speech and noise spectra estimates, and the noisy input spectra are multiplied by the computed gains to produce the noise suppressed output spectra. When implementing the SS technique, we should determine the relevant parameters so as to compromise between the residual noise and the speech distortion i.e., we can reduce the level of residual noise at the cost of increase in speech distortion or vice versa. As a result, it is possible to decrease the number of insertion errors especially during the non-speech periods if we design the SS algorithm to have a lower level of residual noise.

In this paper, we propose a new feature compensation technique called the soft decision IMM (SDIMM) approach in which both the IMM and SS methods are simultaneously applied. SDIMM combines the clean speech estimates provided by the IMM and SS modules depending on the speech absence probability (SAP) which is a byproduct of a usual speech enhancement technique [5]. From a number of speech recognition experiments on AURORA2 database, we can find that the SDIMM approach improves the performance of the original IMM technique.

2. IMM and SS

In this section, we briefly review the IMM and SS techniques used for feature compensation in noisy environments. Interested readers are referred to [3] and [5] for a detail of the two approaches.

2.1. Interacting Multiple Model (IMM)

The IMM approach is usually applied in the log spectral domain. If we let $\mathbf{x} = [x_1, x_2, \cdots, x_D]^T$, $\mathbf{n} = [n_1, n_2, \cdots, n_D]^T$ and $\mathbf{z} = [z_1, z_2, \cdots, z_D]^T$ denote the log spectra of the clean speech, added noise and noisy speech, respectively, then their relation is described as

$$z_i = x_i + \log\left[1 + \exp\left(n_i - x_i\right)\right] \quad , \text{ for } i = 1, 2, \cdots, D \, . \tag{1}$$

In the IMM technique, the probability density function (pdf) for the clean speech is assumed to be a Gaussian mixture distribution given by

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

where M is the total number of mixture components and $p(k)$, μ_k, and Σ_k represent the given weight, mean and covariance of the kth Gaussian, respectively. The parameters $\{p(k), \mu_k, \Sigma_k\}$ are obtained by training on a set of clean speech database, which characterizes the IMM approach as a *task-dependent* technique. In order to make the nonlinear relationship among $\mathbf{z}$, $\mathbf{n}$ and $\mathbf{x}$ a more tractable one, (1) is approximated by a piecewise linear model given by

$$\mathbf{z} \quad = \quad A_k\mathbf{x} + B_k\mathbf{n} + C_k \tag{3}$$

if $\mathbf{x}$ is assumed to have come from the kth mixture component. The coefficient matrices $\{A_k, B_k, C_k\}$ shown in (3) are obtained by the statistical linear approximation (SLA) algorithm which is based on Taylor series expansion of a nonlinear function [2].

A major advantage of the IMM algorithm is that it incorporates a slowly evolving environment model to be able to track the time-varying noise characteristic. The evolving environment model in conjunction with the piecewise linear observation model described in (3) enables us to form a state space model for each mixture component as follows:

$$\begin{aligned} \mathbf{n}_t &= \mathbf{n}_{t-1} + \mathbf{w}_t \\ \mathbf{z}_t &= A_k\mathbf{x}_t + B_k\mathbf{n}_t + C_k \end{aligned} \tag{4}$$

where $\mathbf{x}_t$, $\mathbf{n}_t$ and $\mathbf{z}_t$ respectively represent the log spectra of the clean speech, noise and noisy speech at time instant t and $\mathbf{w}_t$ is a zero-mean Gaussian noise accounting for the evolving nature of the background noise. In (4), the noise feature vector $\mathbf{n}_t$ is treated as the state variable and it is distributed according to a Gaussian pdf $\mathcal{N}(\mathbf{n}_t; \mu_{\mathbf{n}}(t), \Sigma_{\mathbf{n}}(t))$ with $\mu_{\mathbf{n}}(t)$ and $\Sigma_{\mathbf{n}}(t)$ denoting respectively the mean and covariance of the noise log spectrum at time t.

The parameters $\{\mu_{\mathbf{n}}(t), \Sigma_{\mathbf{n}}(t)\}$ concerned with the background noise are sequentially estimated by the IMM algorithm which consists of several steps summarized in the following [3]:

- *Mixing step*: the parameter estimates of the noise obtained from each mixture component in the previous time are combined together to produce a single noise estimate 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 mixture component is updated.
- *Output generation step*: the noise parameter estimates are generated by combining the estimates of all the mixture components. Here, this step is the same to the *Mixing step* mentioned above.

After estimating the noise parameters $\lambda_{\mathbf{n}}(t) = \{\mu_{\mathbf{n}}(t), \Sigma_{\mathbf{n}}(t)\}$ for each time t, the clean speech estimate is computed according to the minimum mean square error (MMSE) criterion such that

$$\hat{\mathbf{x}}_t \quad = \quad E\left[\mathbf{x}_t | \mathbf{z}_t, \hat{\lambda}_{\mathbf{n}}(t)\right] \tag{5}$$

where $\hat{\lambda}_{\mathbf{n}}(t)$ is the estimate for $\lambda_{\mathbf{n}}(t)$ and $E[\cdot]$ means the expectation operation.

2.2. Spectral Subtraction (SS)

In contrast to IMM, the SS approach is performed in the linear spectral domain. Let $\mathbf{Z} = [Z_1, Z_2, \cdots, Z_D]^T$ denote the spectrum of the noisy speech signal with Z_i being the ith spectral component. Two hypotheses H_0 and H_1, which respectively indicate speech absence and presence, are described as follows:

$$\begin{aligned} H_0 &: \mathbf{Z} = \mathbf{N} \\ H_1 &: \mathbf{Z} = \mathbf{X} + \mathbf{N} \end{aligned} \tag{6}$$

where $\mathbf{N} = [N_1, N_2, \cdots, N_D]^T$ represents the spectrum of the added noise that is uncorrelated with the clean speech spectrum $\mathbf{X} = [X_1, X_2, \cdots, X_D]^T$. In the conventional SS technique, it is assumed that $\mathbf{X}$ and $\mathbf{N}$ are characterized by separate zero-mean complex Gaussian distributions though other statistical models are also possible. The estimate for the clean speech spectrum is obtained by

$$\hat{X}_i \quad = \quad \tilde{G}_i \cdot Z_i \quad , \text{ for } i = 1, 2, \cdots, D \tag{7}$$

where $\hat{X}_i$ represents the enhanced ith spectral component and $\tilde{G}_i$ indicates the applied gain.

Among a variety of ways to formulate the spectral gains $\left\{\tilde{G}_i\right\}$, the noise suppression rule proposed by Ephraim and Malah [4], which we call the EMSR method, is the most well-known approach due to its superiority in reducing musical noise phenomena after enhancement. According to the EMSR rule the gain appears as a function of two variables such that

$$\tilde{G}_i \quad = \quad G(\eta_i, \gamma_i) \tag{8}$$

in which η_i and γ_i are referred to the a priori and the a posteriori signal-to-noise ratios (SNR's), respectively. The gain function $G(\cdot,\cdot)$ in (8) is given by

$$G(\eta, \gamma) \quad = \quad \frac{\sqrt{\pi}}{2}\sqrt{\frac{\eta}{\gamma(1+\eta)}} \times M\left[\frac{\gamma\eta}{1+\eta}\right] \tag{9}$$

where

$$M[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right] \tag{10}$$

with I_0 and I_1 being the modified Bessel functions of zero and first order, respectively. In the EMSR method, η_i and γ_i play the key role and their estimates are efficiently updated for each time by means of the decision-directed approach [4].

Given a noisy spectrum $\mathbf{Z}$, we can compute the SAP, $p(H_0|\mathbf{Z})$ such that

$$p(H_0|\mathbf{Z}) \quad = \quad \frac{p(\mathbf{Z}|H_0)\,p(H_0)}{p(\mathbf{Z}|H_0)\,p(H_0) + p(\mathbf{Z}|H_1)\,p(H_1)} \tag{11}$$

where $p(H_0)$ ($= 1 - p(H_1)$) is the a priori probability for speech absence. The likelihoods $\{p(\mathbf{Z}|H_0), p(\mathbf{Z}|H_1)\}$ are specified based on the assumed statistical models for the speech and noise spectra distributions. If the clean speech spectrum $\mathbf{X}$ and the noise spectrum $\mathbf{N}$ are characterized by two independent zero-mean complex Gaussian pdfs, we have

$$p(H_0|\mathbf{Z}) \quad = \quad \frac{1}{1 + q\prod_{i=1}^{D}\Lambda_i(Z_i)} \tag{12}$$

in which q is the ratio defined by

$$q \quad = \quad \frac{p(H_1)}{p(H_0)} \tag{13}$$

and $\Lambda_i(Z_i)$ is the likelihood ratio computed for the ith spectral component as given by

$$\Lambda_i(Z_i) = \frac{1}{1+\eta_i}\exp\left[\frac{\gamma_i\eta_i}{1+\eta_i}\right]. \qquad (14)$$

The SAP is usually applied to modify the gain function shown in (8), and it also becomes an important parameter in the SDIMM approach which will be presented in the following section.

Even though the SS method is performed in the linear spectral domain, we can implement an almost equivalent algorithm directly in the log spectral domain. Let $\hat{\mathbf{X}} = \left[\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_D\right]^T$ be an estimated clean speech spectrum and $\hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_D]^T$ be the corresponding log spectrum. Then,

$$\begin{aligned} \hat{x}_i &= \log\left|\hat{X}_i\right|^2 \\ &= 2\log\tilde{G}_i + \log|Z_i|^2 \\ &= 2\log\tilde{G}_i + z_i \qquad (15) \end{aligned}$$

in which z_i denotes the ith log spectral component of the noisy signal. This kind of approach was proposed in [6] under the name of cepstrum subtraction.

3. SOFT DECISION IMM

Despite the great improvement in recognition performance, one of the drawbacks of the IMM approach is that it causes many undesirable insertion errors in the recognition results. Inappropriate model approximation given by (3) during the non-speech periods is considered to be responsible for this phenomenon. Moreover, there is no way to detect active speech regions in the IMM algorithm. On the other hand, the SS approach takes into account both the models for speech absence and presence simultaneously, and most of its sub-modules depend on the SAP. For that reason, the background noise which exists in the non-speech periods can be more effectively suppressed resulting in less insertion errors. But, this is achieved at the cost of some speech distortion, which may give rise to a number of substitution errors.

In Fig. 1, we give an example in which the energy contours of a noisy speech after feature compensation are displayed and compared with that of the corresponding clean speech. As mentioned above, the energy contour obtained by the IMM algorithm during the non-speech periods makes a lot of erroneous estimates even though it can track the clean speech energy quite well in the active speech regions. In contrast, the SS approach suppresses the noise level during the non-speech periods so as to make the energy contour fit to that of the target clean speech while it causes considerable amount of distortions in the active speech spectra.

In order to reduce the number of insertion errors while maintaining the good performance of the IMM approach in active speech regions, we propose a new feature compensation technique which is called the SDIMM algorithm. A block diagram of the SDIMM algorithm is shown in Fig. 2 where we can find that both the conventional IMM and SS modules are implemented in parallel. The basic idea of the SDIMM algorithm is to combine the two separate estimates provided by the IMM and SS modules. Furthermore, contribution of each module when combining the two estimates is controlled in accordance with the SAP which is computed in the SS module. Let $\hat{\mathbf{x}}_{\text{IMM}}$ be the clean speech feature estimate provided by the IMM algorithm, and $\hat{\mathbf{x}}_{\text{SS}}$ be the estimate obtained based on the SS technique as given by (15). Then, the SDIMM algorithm combines the two estimates as follows:

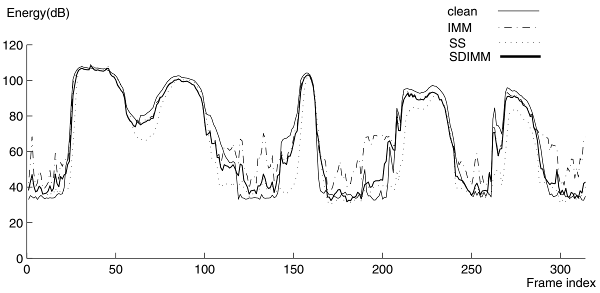

Figure 1: Energy contours of the clean and compensated speech.

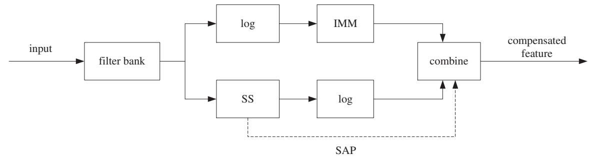

Figure 2: Block diagram of the SDIMM algorithm.

$$\hat{\mathbf{x}} = (1 - p(H_0|\mathbf{Z}))\,\hat{\mathbf{x}}_{\text{IMM}} + p(H_0|\mathbf{Z})\,\hat{\mathbf{x}}_{\text{SS}} \qquad (16)$$

where $p(H_0|\mathbf{Z})$ indicates the SAP with $\mathbf{Z}$ being the given noisy speech spectrum. The IMM module dominates in the feature compensation operation if $p(H_0|\mathbf{Z}) \approx 0$ i.e., within the active speech periods, and the SS module becomes more important as the SAP grows.

4. EXPERIMENTAL RESULTS

Performance of the proposed SDIMM algorithm was evaluated on the AURORA2 database which consists of the TI-DIGITS data downsampled to 8 kHz [8], [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 [10]. Speech features for recognition consisted of twelve cepstral coefficients derived from 23 mel-spaced triangular filters and log energy, and these thirteen parameters were augmented with the corresponding delta and acceleration coefficients. All the HMM's were trained with the clean speech database (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). In the IMM algorithm, clean speech log spectra were modeled by a mixture of 128 Gaussian distributions with diagonal covariance matrices. A gain $\tilde{G}_i$ was computed for each frequency band of the mel-spaced filter bank, and it was applied to derive the correction factor as given by (15) in the SS algorithm.

The recognition results obtained form the AURORA2 task in clean training condition are shown in Table I. From the results, we can easily observe that both the IMM and SDIMM outperformed the SS approach in most of the tested conditions. In addition, the SDIMM approach improved the performance of the IMM algorithm up to 14.26 % which was mainly achieved by reducing the number of insertion errors. These results confirm us that the combined use of the *task-dependent* and *task-independent* approaches based on a soft decision for speech absence is very effective in feature compensation.

	set A	set B	set C	Average
Baseline	61.34	55.75	66.14	60.06
SS	76.23 (37.83)	71.54 (42.85)	74.67 (30.86)	74.04 (38.44)
IMM	80.69 (41.47)	81.35 (58.92)	76.23 (27.28)	80.06 (45.61)
SDIMM	82.89 (50.72)	82.93 (64.42)	78.25 (37.46)	82.17 (53.55)

Table 1: word accuracies (%) over Aurora2 database.
(Relative improvement compared to the baseline system)

5. Conclusions

In this paper, we have analyzed the conventional IMM and SS approaches applied to feature compensation for robust speech recognition in adverse environments. Based on the analysis, we have proposed a new approach called the SDIMM algorithm which combines the two techniques depending on a soft decision for speech activity. The SDIMM algorithm has been found to improve the recognition performance not only in the active speech regions but also during the non-speech periods.

6. References

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

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

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

[4] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean squre error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 12, pp. 1109-1121, Dec. 1984.

[5] N. S. Kim and J. -H. Chang, “Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp.108-110, May 2000.

[6] H. K. Kim, R. C. Rose and H. G. Kang, “Acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments," *Proc. of EuroSpeech*, pp. 421-424, 2001.

[7] 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*, vol.1, pp.409-412, May. 2002.

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

[9] 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.

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