二段雑音スペクトルの推定と回帰による車内音声認識

李　衛鋒[†]　伊藤　克亘[††]　武田　一哉[††]　板倉　文忠[†††]

[†] 名古屋大学工学研究科
[††] 名古屋大学情報科学研究科
[†††] 名城大学情報工学科

あらまし　走行中の車内のような騒々しい環境においても，精度の高い音声認識技術が望まれている．本稿では，二段雑音スペクトルの推定という手法を提案され，一つの遠隔マイクロホンで収録した音声データを基に，非線形回帰を行うことで，車内での音声認識精度の向上を目指した．12 車内走行条件の音声認識実験によって，もとの遠隔マイクロホンに比べて相対ワード認識誤りを 65%の程度で減少できる結果が得られた．

キーワード　非線形回帰 多層パーセプトロン　音声強調　音声認識

Two-stage Noise Spectra Estimation and Regression based In-car Speech Recognition using Single Distant Microphone

Weifeng LI[†], Katunobu ITOU[††], Kazuya TAKEDA[††], and Fumitada ITAKURA[†††]

[†] Graduate School of Engineering, Nagoya University,Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
[††] Graduate School of Information Science, Nagoya University,Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
[†††] Faculty of Science and Technology, Meijo University, Meijo University, Nagoya, 468-8502 Japan.

Abstract In this paper, we present a two-stage noise spectra estimation approach. After the first-stage noise estimation using the improved minima controlled recursive averaging (IMCRA) method, the second-stage noise estimation is performed by employing a maximum a posteriori (MAP) noise amplitude estimator. We also develop a regression-based speech enhance system by approximating the clean speech with the estimated noise and original noisy speech. Evaluation experiments show that the proposed two-stage noise estimation method results in lower estimation error for all test noise types. Compared to original noisy speech, the proposed regression-based approach obtains an average relative word error rate (WER) reduction of 65% in our isolated word recognition experiments conducted in 12 real car environments.

Key words maximum a posteriori (MAP) estimation, spectral subtraction, speech enhancement, multi-layer perceptron, speech recognition

1. Introduction

Noise spectra estimation plays a fundamental role in speech enhancement and speech recognition. Conventional noise estimation methods, which are based on the explicit detection of voice activity, can be difficult in the case of varying background noise or if the signal-to-noise (SNR) is low. In [1], a number of methods which do not need any explicit voice activity detectors (VADs), such as energy clustering, Hirsch histograms, low energy envelope tracking, and so on, are excellently summarized. With picking a quantile value rather than the minima value, quantile based method [2] can be viewed as a generalization of the minimum statistics (MS) approach [3]. More recently, Cohen proposed an improved minima controlled recursive averaging (IMCRA) approach [4] which involved the use of minimum statistics and speech presence probability. On the other hand, once the estimated noise spectra are obtained, one can employ an enhancement filter to estimate the spectral amplitude (or component) of a speech signal in the second stage, by assuming an *ad hoc* statistical model for speech and noise [5] [6]. In this paper, we estimate the spectral amplitude (or component) of the noise signal in a similar manner to that used in speech spectral estimation in the second stage. Therefore, a two-stage noise spectra estimation is developed. In light of the statistical information for short-time spectral amplitude (or component), the second-stage noise estimation can be expected to yield a further improvement of estimation performance. In this paper, specifically, we develop a second-stage *maximum a posteriori* (MAP) noise amplitude estimator based on first-stage IMCRA noise estimation. However, the methods used in the first stage and second stage are not limited, and can be extended to other types of first-stage and second-stage noise estimators. The finally estimated noise spectra can be further integrated into a speech enhance system.

Among a variety of speech enhancement methods, *spectral subtraction* (SS) [7] based method and *short-time spectral estimation* (STSE) based method are commonly applied. Most of SS based methods make assumptions about the uncorrelation of the speech and noise spectra, allowing for simple linear subtraction of the estimated noise spectra. Although scaling factors for emphasis or deemphasis of the estimated noise have been proposed (e.g.,[8]) to reduce the musical tone effects, the specification of the scaling factors is usually done experimentally and is never statistical. The STSE based methods can lead to a nonlinear spectral estimator by introducing a *a priori* SNR, however, they requres the assumptions about an *ad hoc* statistical model for speech and noise [5] [9] [6].

To realease the assumputions that SS based and STSE based methods require, some nonlinear estimators have been implemented by through look up tables [10], curve fitting [11] and neural networks [12] [13] [14] [15]. The approach described in this paper uses neural networks to approximate the log spectral of clean speech with the inputs of the log spectra of the noisy speech and estimated noise. The proposed method differ from [12] [13] [14] in that it is a minimum mean square error (MMSE) estimator in the log spectral domain, since MMSE criterion in the log domain is more consistent with the human auditory system and distance metrics used in speech recognition system [16]. Although MMSE estimators in the log spectral domain are employed in [16] [15], the proposed method discriminate itself against them in that we do not make the assumption about the addition of log power spectra, which may not hold, and that we do not need to estimate the mean and variance of the log speech spectra. While the previous works are usually evaluated on the simulated noisy data, i.e., by artificially adding the noise to the clean speech, the proposed approach is implemented using real stereo data and deals with more general conditions, e.g., channel distortion.

To realease the assumputions that SS based and STSE based methods require, some nonlinear estimators have been implemented by through look up tables [10], curve fitting [11] and neural networks [12] [13] [14] [15]. The approach described in this paper uses neural networks to approximate the log spectral of clean speech with the inputs of the log spectra of the noisy speech and estimated noise. While other neural network based enhancement or compensation methods are implemented in time domain [12] or in cepstrum domain [13] [14], the proposed method is a minimum mean square error (MMSE) estimator in the log spectral domain, since MMSE criterion in the log domain is more consistent with the human auditory system and distance metrics used in speech recognition system [16]. Although MMSE estimators in the log spectral domain are also employed in [16] [15], the proposed method differs from them in that we do not make the assumption about the addition of log power spectra, which may not hold, and in that we do not need to estimate the mean and variance of the log speech spectra.

The organization of this paper is as follows: In Section 2, we present the proposed algorithms including a noise amplitude estimator and the regression method. In Section 3, we evaluate the proposed two-stage noise estimation method. In Section 4, the regression-based in-car speech recognition experiments are described. In Section 5, we summarize this paper.

2. Algorithms

2.1 MAP noise amplitude estimator

We assume that the noisy signal $x(i)$ is given by $s(i)+n(i)$, where $s(i)$ is the clean speech signal which is assumed to be independent of the additive noise $n(i)$. By using short-time Discrete Fourier transform (DFT), in the time-frequency domain we have

$$X(k,l) = S(k,l) + N(k,l),$$

where

$$\begin{aligned} X(k,l) &= R(k,l)\exp\{j\varphi_x(k,l)\}, \\ S(k,l) &= A(k,l)\exp\{j\varphi_s(k,l)\}, \\ N(k,l) &= B(k,l)\exp\{j\varphi_n(k,l)\}, \end{aligned}$$

with the frequency bin index k and the frame index l. We will drop both the frequency bin index k and the frame index l in this subsection, for compactness.

The MAP noise amplitude estimator is given by

$$\hat{B} = \text{argmax } p(R|B)p(B), \tag{1}$$

where $p(\cdot)$ denotes a probability density function (pdf). Let us assume complex Gaussian models for noise and speech spectral components with variances $\lambda_n = E\{|N|^2\}$ and

$\lambda_s = E\{|S|^2\}$, respectively, where $E\{\cdot\}$ denotes the expectation operator, and the variances of their real and imaginary parts are $\lambda_n/2$ and $\lambda_s/2$ respectively. We then have a Rician likelihood $p(R|B)$ and a Rayleigh prior $p(B)$ as

$$p(B) = \frac{2B}{\lambda_n}\exp(-\frac{B^2}{\lambda_n}); \qquad (2)$$

$$p(R|B) = \frac{2R}{\lambda_s}\exp(-\frac{B^2+R^2}{\lambda_s})I_0(\frac{2RB}{\lambda_s}), \qquad (3)$$

where $I_0(z) = \frac{1}{2\pi}\int_0^{2\pi}\exp(z\cos\theta)d\theta$ is the 0-order modified Bessel function of first kind. Following [17], the 0-order modified Bessel function of first kind can be approximated as $I_0(z) \approx e^z/\sqrt{2\pi z}$. For obtaining the noise amplitude estimator, the requirement that the gradient of $\log[p(R|B)p(B)]$ with respect to B vanishes yields

$$2(\frac{1}{\lambda_n}+\frac{1}{\lambda_s})A - \frac{2R}{\lambda_s} - \frac{1}{2B} = 0. \qquad (4)$$

Therefore, the gain function for the noise amplitude estimator can be obtained as

$$G_N = \frac{\hat{B}}{R} = \frac{1}{2(1+\xi)} + \sqrt{\left(\frac{1}{2(1+\xi)}\right)^2 + \frac{1}{4\gamma(1+\frac{1}{\xi})}}, (5)$$

where the *a priori* and *a posteriori* SNRs are defined as $\xi = \lambda_s/\lambda_n$ and $\gamma = R^2/\lambda_n$ respectively [5].

2.2 Regression based enhancement

In this proposed method, we require the reference clean speech for regression training. Let $\mathbf{S}^{(L)}$, $\mathbf{X}^{(L)}$ and $\mathbf{N}^{(L)}$ denote the log mel-filter-bank (MFB) vector obtained from the reference clean speech, the original noisy speech and the estimated noise, respectively. Let $S^{(L)}(m,l)$, $X^{(L)}(m,l)$ and $N^{(L)}(m,l)$ denote their corresponding elements in the filter bank m and at frame l, i.e.,

$$S^{(L)}(m,l) = \log\sum_k r_k^m|S(k,l)|,$$

$$X^{(L)}(m,l) = \log\sum_k r_k^m|X(k,l)|,$$

$$N^{(L)}(m,l) = \log\sum_k r_k^m|N(k,l)|,$$

where r_k^m denotes the weights of the filter bank m. Let $\hat{\mathbf{S}}^{(L)}$ denote the estimated log MFB vector obtained from $\mathbf{X}^{(L)}$ and $\mathbf{N}^{(L)}$. Each element of the log MFB vector of the reference clean speech is approximated independently by employing multi-layer perceptron regression method, where the network with one hidden layer composed of 8 neurons is used, i.e.,

$$\hat{S}^{(L)}(m,l) = b_m + \sum_{p=1}^{8}\left(w_{m,p}\tanh\left(b_{m,p} + w_{m,p}^{x}X^{(L)}(m,l) + w_{m,p}^{n}N^{(L)}(m,l)\right)\right), \qquad (6)$$

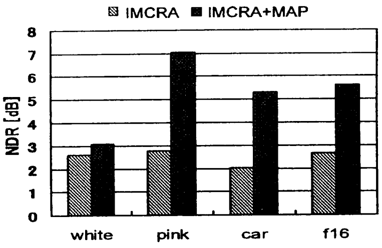

図 1 Averaged NDR values for the IMCRA and the two-stage IMCRA+MAP noise estimators.

where tanh(·) is the tangent hyperbolic activation function. The parameters $\Theta = \{b_m, w_{m,p}, w_{m,p}^x, w_{m,p}^n, b_{m,p}\}$ are found by minimizing the mean squared error:

$$\mathcal{E}(m) = \sum_{l=1}^{L}[S^{(L)}(m,l) - \hat{S}^{(L)}(m,l)]^2, \qquad (7)$$

through the back-propagation algorithm [18]. Here, L denotes the number of training examples.

Although both the proposed regression-based method and *log-spectra amplitude* (LSA) estimator [9] employ the MMSE cost function in the log domain, the former makes no assumptions regarding the distributions of the spectra of speech and noise and does not require the estimation of *a priori* SNR. In addition, compared to *generalized spectral subtraction* (GSS) [7], it makes no assumptions about the independence of the spectra of speech and noise and can benefit from the regression weights, which are statistically optimized. Although a parametric formulation of GSS have been developed in [19] by using MMSE optimization, it also requires assumptions about the independence of the spectra of speech and noise. To calculate the parameter weights, it requires further assumptions regarding the distributions of the spectra of speech and noise, and the estimation of *a priori* SNR. In our proposed method, the parameter weights are obtained by statistical regression training. Therefore, we do not need to assume the independence of speech and noise. Furthermore, the proposed method does not require the distributions of the spectra of speech and noise, nor *a priori* SNR.

3. Evaluation of noise estimation

The noise signals used in our evaluation are taken from the Noisex92. They include white noise, pink noise, car noise and F16 cockpit noise. The speech signals include 100 phonetically balanced sentences (10 sentences for each of 5 female speakers and 5 male speakers), which are recorded using a close-talking microphone when the car is stopped with the engine running (CIAIR in-car speech corpus [20]). The speech signals are degraded by various types of noise with SNRs in the range [-5, 15] dB. Speech signals are digitized

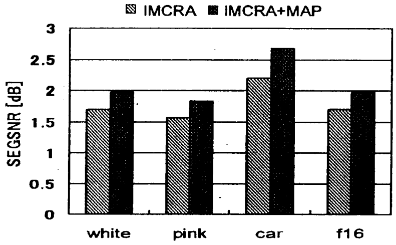

図 2 Averaged segmental SNR improvement for the enhanced speech using the IMCRA and the two-stage IMCRA+MAP noise estimators.

into 16 bits at a sampling frequency of 16 kHz. The spectral analysis is implemented with hamming window of 32 ms (512 samples) and a shift of 16 ms.

To compute the gain function in (5), λ_n is obtained by the IMCRA method [4]. A priori SNR is calculated by the well-known "decision-directed" approach [5]. We compare the noise spectral estimation performance using the noise-to-deviation ratio (NDR), which is defined as

$$\mathrm{NDR}\ [\mathrm{dB}] = 10\log_{10}\sum_{l=1}^{L}\frac{\sum_k[\lambda_n(k,l)]^2}{\sum_k[\lambda_n(k,l)-\hat{\lambda}_n(k,l)]^2}, \quad (8)$$

where λ_n and $\hat{\lambda}_n$ denote the reference noise power spectral and the noise power spectral as estimated by the tested method, and L is the number of frames in the analyzed signal. Fig. 1 presents the results of NDR values averaged over [-5, 15] dB by the IMCRA and the proposed IMCRA+MAP estimators for various noise types. It shows that the latter estimator obtains significantly higher NDR values.

We also examine the performance of the proposed estimation method when integrated into a speech enhancement system. We applied a MAP speech amplitude estimator for speech enhancement, in which the gain function can be obtained in a similar manner to the MAP noise amplitude and is given as

$$G_S = \frac{\hat{A}(k,l)}{R(k,l)} = \frac{1}{2(1+\frac{1}{\xi})} + \sqrt{\left(\frac{1}{2(1+\frac{1}{\xi})}\right)^2 + \frac{1}{4\gamma(1+\frac{1}{\xi})}}. \quad (9)$$

Note that the difference between Equation (5) and Equation (9). We measure the resulting enhanced speech using segmental SNR defined as

$$\mathrm{SegSNR}\ [\mathrm{dB}] = \frac{10}{L}\sum_{l=1}^{L}\log_{10}\frac{\sum_j[s(l,j)]^2}{\sum_j[s(l,j)-\hat{s}(l,j)]^2} \quad (10)$$

where s and $\hat{s}$ denote the reference clean speech and enhanced speech respectively. L is the number of frames in one utterance. Fig. 2 summarizes the results of the segmental SNR improvement for various noise types (averaged over [-5, 15] dB for each type). The enhanced speech obtained by using the proposed IMCRA+MAP noise estimators consistently yields a higher improvement in the segmental SNR for all noise types.

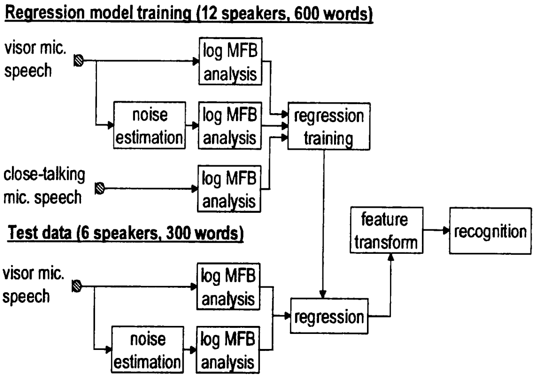

図 3 Diagram of regression-based speech recognition.

4. In-car speech recognition experiments

The speech data used is from CIAIR in-car speech corpus [20]. The speech captured by a microphone at the visor position is used for recognition experiments. The speech collected at a close-talking microphone (by wearing a headset) is referred to as reference speech. Speech signals are digitized into 16 bits at a sampling frequency of 16 kHz. For spectral analysis, 24-channel mel-filter-bank (MFB) analysis is performed on 25 millisecond-long windowed speech, with a frame shift of 10 milliseconds. Spectral components lower than 250 Hz are filtered out because the spectra of the engine noise are concentrated in the low-frequency region. Then log MFB parameters are estimated. The estimated log MFB vectors are transformed into CMN-MFCC vectors using Discrete Cosine Transformation (DCT), and then the time derivatives are calculated. The final feature vectors used in the recognition system consist of 12 CMN-MFCCs + 12 Δ CMN-MFCCs + Δ log energy.

We performed isolated word recognition experiments on the 50 word sets under 12 real car driving conditions (3 driving environments × 4 in-car states as listed in TABLE 1).

表 1 12 driving conditions

driving environment	idling city expressway
in-car state	normal air-conditioner (AC) on low level air-conditioner (AC) on high level window (near the driver) open

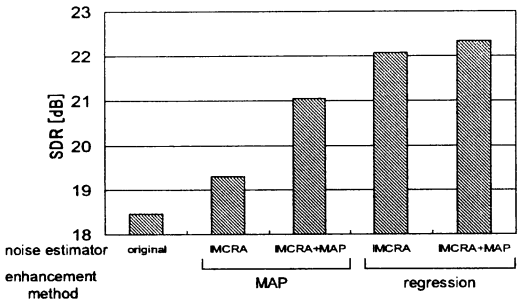

図 4 Averaged SDR values defined in Equation (11) for different speech.

Fig. 3 shows a block diagram of the regression-based speech enhancement system for a particular driving condition. For each driving condition, the data uttered by 12 speakers is used for learning the regression weights and the remaining words uttered by 6 speakers (3 male and 3 female) are used for open testing. Two versions of 1,000-state triphone Hidden Markov Modes (HMM) with 32 Gaussian mixtures per state trained with a total of 7,000 phonetically balanced sentences (3,600 were collected in the idling-normal condition and 3,400 were collected while driving the DCV on the streets near Nagoya university (city-normal condition)), are used for acoustical models. One, by the name of "CloseTalking-HMM", is trained with 7,000 phonetically balanced sentences collected at the close-talking microphone, and the other, by the name of "Visor-HMM", is trained with the sentences collected at the visor microphone.

For comparison, a MAP speech amplitude estimator in Equation (9) is also applied, Improved minima controlled recursive averaging (IMCRA) method [4] was used to estimate the noise estimation. The two-stage noise estimator (named "IMCRA+MAP"), in which MAP noise amplitude estimator was employed after the first-stage IMCRA. was also performed for comparison.

We first evaluated the approximation performance of the proposed regression method and two-stage noise estimator, by using the signal-to-deviation ratio (SDR), which is given by

$$\mathrm{SDR}\ [\mathrm{dB}] = 10\log_{10}\frac{\sum_l \sum_m [S^{(L)}(m,l)]^2}{\sum_l \sum_m [S^{(L)}(m,l) - \hat{S}^{(L)}(m,l)]^2}, \quad (11)$$

where $S^{(L)}(m,l)$ and $\hat{S}^{(L)}(m,l)$ denote the reference log MFB element from the close-talking microphone and the estimated log MFB element respectively. L denotes the number of frames during one utterances. The SDR values are averaged over the number of utterances. Fig. 4 shows the SDR values obtained using different methods (averaged over 12 driving conditions). With IMCRA estimator, the regression method yields higher SDR value, which demonstrates the better approximation performance, compared to the enhanced speech using (9). SDR values are further improved by using IMCRA+MAP noise estimators. The regression method using the IMCRA+MAP noise estimator yields the highest SDR, which results in an improvement of approximately 4 dB compared with that of the original speech. These results clearly demonstrate the effectiveness of the proposed regression method and two-stage noise estimator.

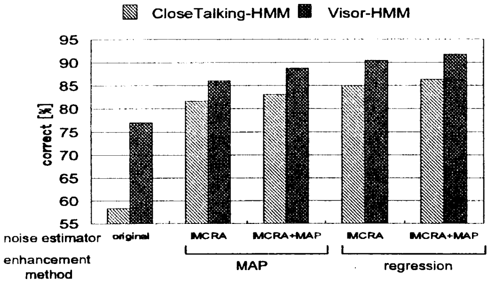

図 5 Averaged word recognition performance for different speech.

Fig. 5 shows the recogntion performance (averaged over 12 driving conditions). With "CloseTalking-HMM" and IMCRA noise estimator, the enhanced speech using (9) provides a significant improvement compared to the original speech. The proposed regression method yields furthermore higher recogntion accuracy, which can be expected from the SDR values in Fig. 4. Using the two-stage IMCRA+MAP noise estimator provides a further improvement. The recognition results with "Visor-HMM" are consistent with the ones with "CloseTalking-HMM" . "Visor-HMM" performs better than "CloseTalking-HMM", in our opinions, in that the mismatch between the training data (idling-normal and city-normal conditions) collected at the visor microphone and the enhanced test data is smaller. The regression method with IMCRA+MAP noise estimator and "Visor-HMM" performs best and achieves an accuracy of 91.7%, an average relative word error rate (WER) reduction of 65% compared to original noisy speech.

5. Summary

In this paper, a two-stage noise spectra estimation approach and a regression-based speech enhancement approach are proposed. The second-stage enhancement-filter-like noise estimation is performed after the first-stage conventional noise estimation. In the proposed regression-based speech enhance system, the log spectra of the clean speech are approximated by using those of the estimated noise and the original noisy speech. Lower estimation errors are obtained by using the proposed two-stage noise estimation method. Use of the regression-based method results in a significant improvement in recognition accuracy.

文　献

[1] C. Ris and S. Dupont, "Assessing local noise level estimation methods: application to noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 141–158, 2001.

[2] V. Stahl; A. Fischer and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Proc. IEEE ICASSP*, 2000, pp. 1875–1878.

[3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[4] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 476–475, 2003.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE ICASSP*, 2002, pp. 253–256.

[7] J. R. Deller; J. G. Proakis and J. H. L. Hansen, *Discrete-time processing of speech signals*, New York: Maxwell Macmillian, 1993.

[8] M. Berouti; R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, 1979, vol. 4, pp. 208–211.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Trans. ASSP*, vol. 32, no. 2, pp. 443–445, 1985.

[10] J.E. Porter and S.F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE ICASSP*, 1984, pp. 18.A.2.1–18.A.2.4.

[11] F. Xie and D. V. Compernolle, "Speech enhancement by nonlinear spectral estimation — a unifying approach," in *Proc. EUROSPEECH*, 1993, pp. 617–620.

[12] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in *Proc. IEEE ICASSP*, 1988, pp. 553–556.

[13] Helge B. D. Sorensen, "A cepstral noise reduction multilayer neural network," in *Proc. IEEE ICASSP*, 1991, pp. 993–996.

[14] K. Ng; H. Gish and J. R. Rohlicek, "Robust mapping of noisy speech parameters for hmm word spotting," in *Proc. IEEE ICASSP*, 1992, pp. 109–112.

[15] F. Xie and D. V. Compernolle, "Speech enhancement by spectral magnitude estimation — a unifying approach," *Speech Communication*, vol. 19, no. 2, pp. 89–104, 1996.

[16] F. Xie and D. V. Compernolle, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Proc. IEEE ICASSP*, 1994, pp. 53–56.

[17] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[18] S. Haykin, *Neural Networks - A Comprehensive Foundation*, Prentice Hall, 1999.

[19] B. Sim; Y. Tong; S. Chang and C. Tan, "A parametric formulation of the generalized spectral subtraction method," *Trans. SAP*, vol. 6, no. 4, pp. 328–337, 1998.

[20] N. Kawaguchi; S. Matsubara; H. Iwa; S. Kajita; K. Takeda; F. Itakura and Y. Inagaki, "Construction of speech corpus in moving car environment," in *Proc. IEEE ICSLP*, 2000, pp. 362–365.