Proc. Int. Conf. Spoken Language Processing (Banff, AB), pp. 515-518, Oct. 1992

SPEECH ENHANCEMENT USING A STATISTICALLY DERIVED FILTER MAPPING

Yan Ming Cheng, Douglas O'Shaughnessy and Peter Kabal

INRS-Telecommunications, Université du Québec
16 Place du Commerce, Nuns' Island, Quebec H3E 1H6 Canada

Abstract

We view the speech enhancement task in two aspects: reduction of the perceptual noise level in degraded speech and reconstruction of the degraded information, which may result in improvement of speech intelligibility. We are also very interested in noise-independent speech enhancement where test noise environments could differ in intensity from those of algorithm development. To this end, we have developed in this paper an algorithm called Noise-Independent Statistical Spectral Mapping (NISSM) to estimate a speech enhancement Wiener filter. NISSM consists of a noise-resistant transformation, which converts noisy speech to a set of noise-resistant features, and a spectral mapping function, which maps the features to autoregressive spectra of clean speech. We will show that the proposed algorithm effectively reduces noise intensity. When the noise intensity of training differs from that of testing, NISSM outperforms significantly a conventional spectral mapping. The algorithm operates frame-by-frame and is designed for real-time application. The noise interference could be stationary or non-stationary white noise with variable intensity.

I. Introduction

Speech enhancement algorithms accept a speech signal which has been degraded by noise, and attempt to reduce the noise and render output speech of improved quality. From the viewpoint of perception, the objectives of speech enhancement are to reduce the noise and to improve the intelligibility of the original speech. Reducing those aspects of the input signal which are due to the noise generally results in a reduction of the perceived noise. Speech enhancement has been investigated for many years, and several successful techniques have been reported (e.g. [1,2,3,4,5]). However, it has been found so far that only a slight improvement of intelligibility can be attained by speech enhancement. Significant intelligibility improvement is a continuing challenge for speech enhancement and is important for practical application. Our lack of precise understanding of speech intelligibility cues has been a major factor in the limited success of improving speech intelligibility of degraded speech signals. One certainty is that recovery of the original speech from the degraded signal should reconstruct (however inefficiently) the carrier of the speech intelligibility cues.

Restricting our view to one-channel speech enhancement, we note several attempts in this direction in recent years. For example, *spectral mapping* [1] maps a noise spectrum directly to a corresponding clean spectrum. A second enhancement method, called *waveform mapping* [2], uses an artificial neural network to map a noisy speech waveform to a clean one. Thirdly, *Hidden Markov Models* (HMM) [3] can also be used to realize a mapping function to enhance speech. An HMM maps a sequence of noisy spectra, instead of a single spectrum or a single waveform vector, to a sequence of clean spectra. Taking advantage of the temporal structure of speech which is noise-resistant and can be captured in the HMM, this approach has practically demonstrated excellent performance.

Mappings have been used as a common means to enhance degraded speech. This approach has plausible reasoning on theoretical grounds, and has demonstrated success on limited applications. However, we have noticed the common weakness that all mapping functions in previous work have depended very much upon the noise level of the speech signals on which they are trained, and that the mapping accuracy decreases when noise level increases. We have also noticed that the above weakness can be minimized by developing a mapping function based on noise-resistant features, if there are any. These insights motivated us to use a mapping technique with noise-resistant features to enhance degraded speech so as to increase, if possible, speech intelligibility. In this paper, we present an algorithm denoted as Noise-Independent Spectral Mapping (NISM) which attempts to construct noise-independent mapping functions derived from speech statistics. In the next section, we give the theoretical description of our algorithm. The issues of practical implementation and experimental justification will be discussed in the third section. The final section gives some concluding remarks on the proposed system.

II. Spectral mapping derived from statistics

The general idea which we use here is introduced in [4], and exploits noise-resistant features of speech which are sufficient to describe the relevant underlying sources of the speech. The difficulty of finding a good mapping depends only on the reliability of the features. A mapping function can be trained from the feature space, instead of from the noisy speech. Then it can be applied to any range of noise level where the noise-resistant property holds.

A. Mapping derivation

Let $\mathbf{s}$ be a sample vector of noisy speech on space $\mathcal{S}$ and let $\mathbf{y}$ be a sample vector of clean speech on space $\mathcal{Y}$, the latter being generated by M independent and mutually exclusive random sources, $\lambda_k, 1 \leq k \leq M$. Assume the noise to be additive and white with

$$\mathbf{s} = \mathbf{y} + \mathbf{n},$$

where $\mathbf{n}$ is a sample vector of noise. The *posterior* pdf (probability density function) of the clean speech given the observed noisy speech vector is

$$p(\mathbf{y}|\mathbf{s}) = \sum_{k=1}^{M} p(\mathbf{y}, \lambda_k|\mathbf{s}) = \sum_{k=1}^{M} p(\lambda_k|\mathbf{s}) p(\mathbf{y}|\lambda_k, \mathbf{s}).$$

Since the noisy signal is independent of the speech signal, we have

$$p(\mathbf{y}|\mathbf{s}) = \sum_{k=1}^{M} p(\lambda_k|\mathbf{s}) p(\mathbf{y}|\lambda_k). \quad (1)$$

The conventional Minimum Mean-Square Estimation (MMSE) of the clean speech is:

$$\hat{\mathbf{y}} = E(\mathbf{y}|\mathbf{s}) = \sum_{k=1}^{M} p(\lambda_k|\mathbf{s})E(\mathbf{y}|\lambda_k) = \sum_{k=1}^{M} p(\lambda_k|\mathbf{s})\mathbf{y}_{\lambda_k}, \quad (2)$$

where $\mathbf{y}_{\lambda_k}$ is a sample vector which corresponds to the expectation feature of the source, λ_k. The mapping functions, $p(\lambda_k|\mathbf{s})$, depend on the noise and tend to resemble a flatter pdf (i.e., more confusable) when the noise level increases.

Let us introduce a vector of noise-resistant features, $\mathbf{x}$, on the space $\mathcal{X}$ and assume also that there are M independent and mutually exclusive random sources, θ_i, which are able to generate all possible $\mathbf{x}$. The *posterior* pdf is

$$\begin{aligned} p(\mathbf{y}|\mathbf{s}) &= \sum_{i=1}^{M}\sum_{k=1}^{M} p(\mathbf{y}, \lambda_k|\theta_i)p(\theta_i|\mathbf{s}) \\ &= \sum_{i=1}^{M}\sum_{k=1}^{M} p(\mathbf{y}|\lambda_k, \theta_i)p(\lambda_k|\theta_i)p(\theta_i|\mathbf{s}). \end{aligned} \quad (3)$$

If we further assume that each source on $\mathcal{X}$ corresponds to one and only one source on $\mathcal{Y}$, i.e., $p(\lambda_k|\theta_i) \triangleq \delta(k-i)$, thus, $p(\mathbf{y}|\lambda_k, \theta_k) \triangleq p(\mathbf{y}|\lambda_k)$ and

$$p(\mathbf{y}|\mathbf{s}) = \sum_{k=1}^{M} p(\mathbf{y}|\lambda_k)p(\theta_k|\mathbf{s}). \quad (4)$$

λ_k and θ_k are chosen by maximizing an *a posteriori* joint pdf, $p(\mathbf{Y}|\mathbf{S}) = p(\mathbf{y}_1, \ldots, \mathbf{y}_t, \ldots, \mathbf{y}_T|\mathbf{s}_1, \ldots, \mathbf{s}_t, \ldots, \mathbf{s}_T)$ over a collection of doublets, $(\mathbf{Y}, \mathbf{S}) = \{(\mathbf{y}_1, \mathbf{x}_1 = f(\mathbf{s}_1)), \ldots, (\mathbf{y}_T, \mathbf{x}_T = f(\mathbf{s}_T))\}$, where $f()$ is a transformation.

$$\begin{aligned} p_{\lambda,\theta}(\mathbf{Y}|\mathbf{S}) &= \max_{\lambda,\theta} p(\mathbf{Y}|\mathbf{S}) = \max_{\lambda,\theta} \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{s}_t) \\ &= \max_{\lambda,\theta} \prod_{t=1}^{T}\sum_{k=1}^{M} p(\mathbf{y}_t|\lambda_k)p(\theta_k|\mathbf{s}_t) \\ &= \max_{\lambda,\theta} \sum_{j=1}^{M^T}\prod_{t=1}^{T} p(\mathbf{y}_t|l_{\lambda,j,t})p(l_{\theta,j,t}|\mathbf{s}_t), \end{aligned} \quad (5)$$

where j is the index of a path consisting of the concatenation of λ_k or θ_k along t; $l_{\lambda,j,t}$ and $l_{\theta,j,t}$ are λ and θ source indices of the j^{th} paths at time t. It has been proven [6] that the above *a posteriori* joint pdf could be asymptotically approximated by the highest *a posteriori* joint pdf on a single path, denoted the Most Likely Path (MLP), among M^T paths, i.e.,

$$p_{\lambda,\theta}(\mathbf{Y}|\mathbf{X}) \approx \max_{\lambda,\theta} \prod_{t=1}^{T} p(\mathbf{y}_t|l_{\lambda,\star,t})p(l_{\theta,\star,t}|\mathbf{s}_t), \quad (6)$$

where $\star$ indicates the MLP. We will see in the next section that eq. (6) can be resolved by partitioning the $\mathcal{X}\times\mathcal{Y}$ space into M cells through the generalized Lloyd algorithm with an assumption that each of the random sources is p^{th}-order autoregressive Gaussian.

Given a set of λ_k and θ_k and a noisy speech sequence, $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_t, \ldots, \mathbf{s}_T\}$, the process of speech enhancement is to find a clean speech sequence, $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_t, \ldots, \hat{\mathbf{y}}_T\}$, which gives the maximum *a posteriori* joint pdf:

$$p_{\lambda,\theta}(\hat{\mathbf{Y}}|\mathbf{S}) = \prod_{t=1}^{T}\sum_{k=1}^{M} p(\mathbf{y}_t|\lambda_k)p(\theta_k|\mathbf{s}_t). \quad (7)$$

Note that the pdfs at time t in eq. (7) are independent of those at any other time. Eq. (7) is equivalent to

$$p_{\lambda,\theta}(\hat{\mathbf{y}}_t|\mathbf{s}_t) = \sum_{k=1}^{M} p(\hat{\mathbf{y}}_t|\lambda_k)p(\theta_k|\mathbf{s}_t), \quad (8)$$

which means that the estimate of a clean speech vector can immediately be obtained through MMSE when a noisy speech vector is supplied; i.e.,

$$\hat{\mathbf{y}}_t = E(\hat{\mathbf{y}}_t|\mathbf{s}_t) = \sum_{k=1}^{M} E(\hat{\mathbf{y}}_t|\lambda_k)p(\theta_k|\mathbf{s}_t) = \sum_{k=1}^{M} \mathbf{y}_{\lambda_k} p(\theta_k|\mathbf{s}_t). \quad (9)$$

B. Transformation of speech to noise-resistant features.

As we have seen, the estimation of clean speech depends strongly upon the assumption that noise-resistant features are practically available. The goodness of the features supports the quality of the mapping. The features we are looking for here are not only noise-resistant but also easy to use with some well-known distortion metrics. The formant frequencies have been proposed as noise-resistant features [4] with a simple Euclidean metric. However, the order of the formants in a time-varying spectrogram of speech is crucial in the Euclidean metric, but formants are difficult to track during many phoneme-to-phoneme transitions. Their reliability is thus questionable for the formant metric. The speech spectral phase has been considered insensitive to noise interference. The lack of a well-established distortion metric using this feature exhibits the difficulty of our objective. Auditory spectra have been suggested to possess noise-resistant properties. A simple processing [5], part of the complete, complex auditory processing of the ear, has recently been proposed to obtain a noise-resistant spectrum. All metrics developed for spectral distortion would be appropriate to apply to this feature. The advantage of this feature motivates us to use it in our enhancement algorithm. The conversion of $\mathbf{s}$ to $\mathbf{x}$ is deterministic, i.e.,

$$\mathbf{x} = f(\mathbf{s}), \quad (10)$$

where $f()$ contains two operations: (1) a Fourier transform of $\mathbf{s}$, and (2) an application of *lateral inhibition* processing to the Fourier amplitude spectrum (algorithm I in [5]). In Fig. 1, we show the noise-resistant properties of $\mathbf{x}$ by its p^{th}-order autoregressive spectra in the presence of noise interference. For comparison, the same spectra of $\mathbf{s}$ have also been plotted. The noise-resistant properties are clearly observed in terms of the relatively invariant spectra of $\mathbf{x}$ as a function of noise strength.

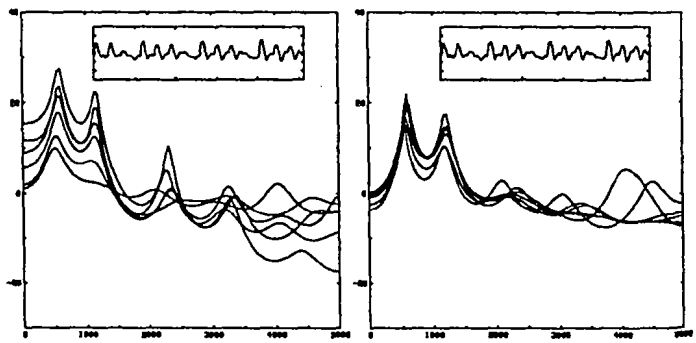

Fig. 1 Spectra as a function of noise intensity. On the left are spectra (amplitude in dB versus frequency in Hz) from a linear prediction analysis (twelfth order) of noisy speech; on the right are those of lateral-inhibition-processed noisy speech. The overlaid plots represent variation in noise intensity: 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB, and ∞ dB. The lowest spectrum corresponds to 0 dB and the highest spectrum to ∞ dB (i.e., no noise). The speech waveform is shown in the inset.

III. Speech enhancement algorithm and experimental results

Our speech enhancement algorithm is generally based on the theory of Wiener filtering. At a given time, a spectrum of clean speech is estimated through MMSE, based on a noisy speech observation. A filter, designed from such a spectrum, is applied to the noisy speech to output an enhanced version of the speech. The actual system consists of three major components: (1) extraction of noise-resistant features, which realizes eq. (10) and outputs a waveform vector corresponding to a zero-phase auditory spectrum; (2) a codebook which contains coefficients of random sources on both $\mathcal{X}$ and $\mathcal{Y}$ spaces; (3) a composite Wiener filter from a bank of weighted filters, each of which are determined by the spectrum of a random source on $\mathcal{Y}$. Fig. 2 summarizes the system overview. Moreover, we investigate in our experiments the idea [7] of improving the enhancement performance iteratively. We will describe each of the components in detail except the $f()$ box, which is adequately described in [5].

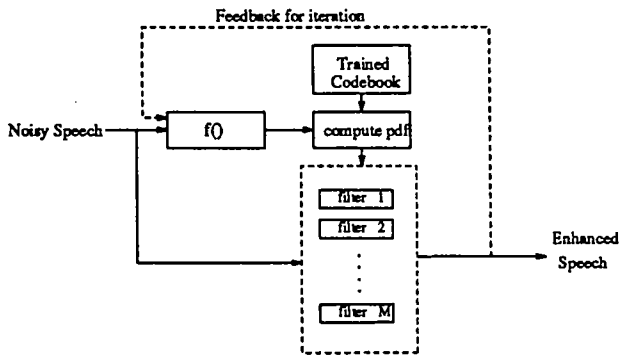

Fig. 2 Block diagram of the speech enhancement system.

A. Speech data

The speech waveforms in our experiments were sampled at 10 kHz, and the speech material was phonetically balanced. At any given time, a frame of 128 speech samples was observed. This frame size is large enough to provide an accurate estimation of the random speech source and is small enough to approximate speech stationarity over the frame. The frame advance was set to 64 samples and a Hamming window was used when necessary. Twelfth-order ($p = 12$) linear prediction analysis was used in the pdf estimation (see later). We collected approximately one hundred minutes of speech as a training set and about thirty minutes as a test set. We trained the mapping function on the noise-resistant features obtained from clean speech, and used it for all ranges of noise level.

B. Training procedure

Instead of maximization in eq. (6), we minimize its negative logarithm:

$$D = -\ln(p_{\lambda,\theta}(\mathbf{Y}|\mathbf{S})) = \min_{\lambda,\theta}\Big\{-\sum_{t=1}^{T}\ln(p(\mathbf{y}_t|l_{\lambda,*,t}))+\ln(p(l_{\theta,*,t}|\mathbf{s}_t))\Big\}. \tag{11}$$

Since each pdf is time-independent, the MLP is thus equal to the concatenation of the "best" sources at each time:

$$D = -\ln(p_{\lambda,\theta}(\mathbf{Y}|\mathbf{S})) = \min_{\lambda,\theta}\Big\{-\sum_{t=1}^{T}\max_{k}\{\ln(p(\mathbf{y}_t|\lambda_k))+\ln(p(\theta_k|\mathbf{s}_t))\}\Big\}. \tag{12}$$

Assuming that all of the random sources on both $\mathcal{X}$ and $\mathcal{Y}$ are p^{th}-order autoregressive Gaussian sources ($p = 12$), we have [8]

$$p(\mathbf{u}|\mathbf{v}) = e^{-\frac{N}{2}d(\mathbf{u},\mathbf{v})} \tag{13.a}$$

$$d(\mathbf{u},\mathbf{v}) = \frac{\mathbf{r}_u'\mathbf{r}_{a_v}}{\sigma_u^2} + \ln(2\pi\sigma_u^2), \tag{13.b}$$

where vector $\mathbf{r}_u$ contains $p+1$ values of the autocorrelation function of process $\mathbf{u}$; σ_u is the variance of the p^{th}-order minimum predictive residual of $\mathbf{u}$; vector $\mathbf{r}_{a_v}$ contains the first $p+1$ values of the autocorrelation function of the autoregressive coefficient sequence of process $\mathbf{v}$:

$$D = \min_{\lambda,\theta}\Big\{\sum_{t=1}^{T}\min_{k}\{\frac{N}{2}[d(\mathbf{y}_t,\lambda_k)+d(\theta_k,\mathbf{s}_t)]\}\Big\}. \tag{14}$$

The solution to this equation is very well-known as the Lloyd algorithm (also popularly called the *k-means* method). It partitions the space $\mathcal{Y}\times\mathcal{X}$ into M cells with a composite distortion measure, $d(\mathbf{y}_t,\lambda_k)+d(\theta_k,\mathbf{s}_t)$. An alternative to the Lloyd algorithm called the LBG method [9] is currently used, because of its advantages for algorithm initialization.

B. Enhancement procedure

The speech enhancement procedure can be expressed through Wiener filtering and eq. (9) as follows:

$$\hat{\mathbf{s}}_t = \mathbf{s}_t * \hat{\mathbf{y}}_t = \sum_{k=1}^{M}\mathbf{s}_t * \mathbf{y}_{\lambda_k}p(\theta_k|\mathbf{s}_t) \approx \sum_{k^{\#}=1}^{M^{\#}<M}\mathbf{s}_t * \mathbf{y}_{\lambda_{k^{\#}}}p(\theta_{k^{\#}}|\mathbf{s}_t), \tag{15}$$

where the sign $*$ represents convolution; the sign $^{\#}$ indicates the first $M^{\#}$ largest $p(\theta_k|\mathbf{s}_t)$, which are used in order to reduce the computation load. The pdf $p(\theta_k|\mathbf{s}_t)$ can be evaluated through eq. (13), given $\mathbf{x}_t$. Since $\mathbf{y}_{\lambda_k}$ is the mean vector of the k^{th} autoregressive Gaussian source or the impulse response of the corresponding autoregressive filter, the above convolution can be compactly realized by p^{th}-order autoregression.

C. Experimental results

We assess here two important parameters for the enhancement algorithm: (1) M, a number that determines how many random Gaussian sources are necessary to describe the compressed speech; and (2) $M^{\#}$, a number that determines how many filters are sufficient to produce accurate enhanced speech given noisy speech. We apply the *Itakura-Saito* (IS) distortion [10] and rms (root-mean-square) log spectra between the enhanced spectra and the corresponding clean spectra as a criterion to test performance. The IS-distortion is more sensitive to spectral peaks than to spectral valleys; rms log spectra treat distortion in the frequency domain equally. We first investigate the enhancement scheme without iterative improvement, i.e., an open loop scheme.

In Fig. 3, we show IS-distortion and rms log spectra for the test set, for SNR = 0 dB and SNR = 10 dB, when M varies from $2^7 = 128$ to 2^{10}=1024 and when $M^{\#}$ varies from $2^0 = 1$ to $2^7 = 128$. We have seen that an increase in M from 128 to 1024 decreases the distortion when $M^{\#}$ is one, i.e., our system acts as a vector quantizer (VQ). An increase in $M^{\#}$ from 1 to 128 decreases significantly the distortion for any M. However, when $M^{\#}$ is large (e.g., 128), the increase in M also increases the distortion unfortunately. The explanation may be that the distortion depends upon the ratio of $\frac{M^{\#}}{M}$. The larger M is, the lower the ratio is. These results suggest that the spectra of clean speech could be effectively reconstructed by contributions of all random sources and that the VQ approach, which takes only the contribution of the one best source, is very sub-optimal.

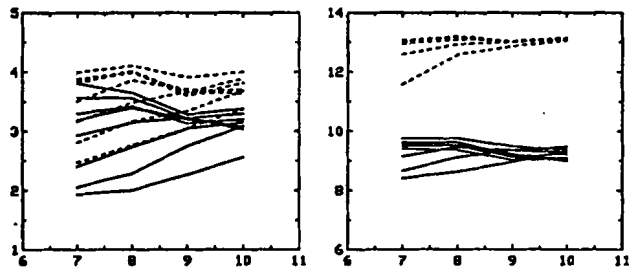

Fig. 3 Developed spectral mapping (NISM) performance, using IS-distortion (left) and rms log spectra (right) as a function of the number of sources, M, and number of filters, $M^{\#}$. Horizontal axes indicate M in bits. Vertical axes are distortion (in an abitrary scale). Dashed lines are used for SNR = 0 dB and solid lines for SNR = 10 dB. The lines from top to bottom correspond to $M^{\#} = 2^0, 2^1, 2^2, \ldots, 2^7$.

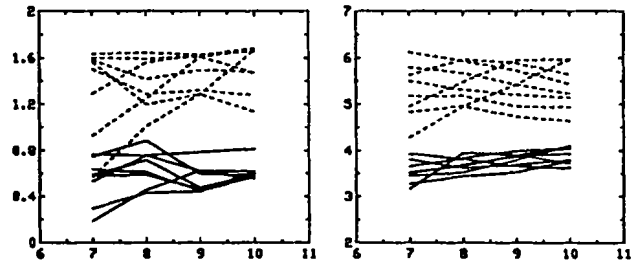

Fig. 4 Increments of the distortion as a function of the increase of noise intensity from SNR 10 dB to SNR = 0 dB. Dashed lines show the increments for the conventional spectral mapping; solid lines are used for the noise-independent spectral mapping. Horizontal axes are M in bits. Vertical axes show distortion in an abitrary scale. The left panel shows IS-distortion and the right shows rms log spectra.

In order to demonstrate the effectiveness of noise-resistant features, we compared the noise-independent spectral mapping with the conventional spectral mapping based on eq. (2), as far as resisting noise interference. We repeated the above experiments for the conventional spectral mapping, and examined the increases in distortion for both a conventional mapping and the noise-independent spectral mapping when SNR is decreased from 10 dB to 0 dB. Fig. 4 plots such increases. The increases in IS-distortion for the conventional spectral mapping are about 2 times as high as that for the noise-independent spectral mapping and the increases with rms log spectra were about 1.5 times as high. Therefore, the noise-resistance in the noise-independent spectralmapping is very effective and the mapping confusion is less severe than that in a conventional spectral mapping.

We have experimented with a closed-loop enhancement scheme, i.e., iteratively improving an estimate of clean speech, based on the noisy speech. At each iteration, the enhanced speech is fed back as noisy speech to compute again all the pdf's, then obtaining a new estimate. Beyond eight iterations, we had not perceived a significant reduction of the distortion. However, when viewing spectrograms, the noise level was slightly reduced compared with an open-loop scheme.

IV. Concluding Remarks

We have developed in this paper a statistical mapping of spectra to enhance degraded speech. The crucial factor which makes the present algorithm superior to conventional spectral mapping, in the aspects of mapping confusion and of noise-independent mapping, is the introduction of noise-resistant features. However, an ideal noise-independent mapping was not achieved in our experiments, since the currently-used noise-resistant features are merely an approximation of ideal features. The performance degradation due to a variable noise environment is significantly lower than that made by conventional spectral mappings. Moreover, the impression of noise in the output speech is effectively reduced by our algorithm in the case of a noise environment differing very much from the training environment. Keeping in mind that an ideal noise-independent mapping is never possible, we can say that our preliminary step toward noise-independent speech enhancement has been successful. Comparing our system with HMM-based enhancement [3] on theoretical grounds, the HMM method relies on the noise-resistant advantage of the temporal structure of speech and our technique uses the auditory spectrum of the speech. These two advantages are not contradictory but complementary. Thus a combination of both will gain yet more enhancement power. For instance, taking a local temporal structure of speech in our algorithm, for instance the use of *look-ahead* or *delayed-decision* in the implementation, such as a system could capture both advantages, and yet can be implemented in a real-time system.

Acknowledgments

This work was supported in part by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada and from Canadian Department of Communications.

Reference

1. B.H. Juang and L.R. Rabiner, "Signal restoration by spectral mapping," *Proc. Int. Conf. ASSP*, pp. 2368-2371, 1987.
2. S. Tamura and A. Waibel, "Noise reduction using connectionist Models," *Proc. Int. Conf. ASSP*, New York, pp. 553-556, 1988.
3. Y. Ephraim, D. Malah and B.H. Juang ,"On the application of Hidden Markov Models for enhancing noisy speech," *IEEE Trans. ASSP*, Vol. 37, No.12, Dec. pp. 1846-1856, 1989.
4. D. O'Shaughnessy, "Speech enhancement using vector quantization and a formant distance measure," *Proc. Int. Conf. IEEE ASSP*, New York, pp. 549-552, 1988.
5. Y.M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," *IEEE Trans. ASSP*, Vol. 39, No.9, pp.1943-1954, Sept. 1991.
6. N. Merhav and Y. Ephraim,"Hidden Markov Modeling using a dominant state sequence with application to speech recognition," *Computer, Speech and Language*, No.5, Dec. pp.327-339, 1991.
7. J.S. Lim and A.V. Oppenheim ,"All-pole modeling of degraded speech," *IEEE Trans. ASSP*, Vol. 26, No.2, Feb. pp. 197-206, 1978.
8. F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. ASSP*, Vol. 23, No.1, Feb. pp. 67-72, 1975.
9. A. Buzo, A.H. Gray Jr., R.M. Gray and J.D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. ASSP-28*, No. 5, pp. 562-574, 1980.
10. R.M. Gray, A. Buzo, A.H. Gray Jr. and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. ASSP-28*, No. 4, pp. 367-376, 1980.