A Segment-Based Algorithm of Speech Enhancement for Robust Speech Recognition

Guokang Fu[*] and Ta-Hsin Li[†]

[*]IBM China Research Lab, Shang Di, Beijing 100085, PRC (fugk@cn.ibm.com)

[†] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598-0218 USA (thl@watson.ibm.com)

Abstract

Accurate recognition of speech in noisy environment is still an obstacle for wider application of speech recognition technology. Noise reduction, which is aimed at cleaning the corrupted testing signal to match the ideal training conditions, remain to be an effective approach to improving the accuracy of speech recognition in noisy environment. This paper introduces a new algorithm of noise reduction that combines a tree-based segmentation method with the maximum likelihood estimation to accommodate the nonstationarity of speech while efficiently suppressing the possibly nonstationary noise. Numerical results are obtained from the experiments on an speech recognition system, showing the effectiveness of the proposed algorithm in improving the accuracy of Chinese speech recognition.

I. Introduction

Commercial speech recognition systems (SRS) have been very successful in many applications. But with the advent of mobile devices, more robust SRS are required to meet the customer' s expectation of high performance even in highly adverse situations. Because SRS reply on statistical features of speech estimated from a set of training data, any mismatch of the training data with the actual condition in which the SRS are used can affect the performance of the SRS. In many applications, the difference between the training data and the actual operating conditions cannot be adequately predicted at the time SRS are deployed, so the SRS cannot be trained in advance to suite these conditions. This necessitates the development of adaptive methods that automatically compensate for the unknown operating conditions.

Among many factors that cause the mismatch between the training data and the actual operating conditions, the background noise is of particular concern. In this paper, we focus on this so-called environmental degradation.

Feature-domain modeling and compensation is a widely-used approach to handling the environmental degradation. Examples include the methods of feature subtraction, normalization, and transformation [1]. An alternative approach is the acoustic-domain signal processing, with the aim of reducing the environmental degradation on the speech waveform or its spectrum. This is the approach we take in developing the proposed algorithm. Although speech enhancement is widely used to describe this approach, speech restoration is a more appropriate term because the aim is to remove the degradation rather than enhance the perceptive quality of the signal, degraded or not.

Existing algorithms for speech restoration include the pioneering work of Lim and Oppenheim [2] and the more recent works [3-6]. One major obstacle for speech restoration is the nonstationarity of the speech signal. Frame-by-frame processing is a typical technique of handling the nonstationarity. More elaborate methods [4] [6] require the knowledge of a nonstationary speech model such as the hidden Markov model (HMM).

In this paper, we propose a segment-based algorithm. The degraded speech signal is first partitioned into quasi-stationary segments, where the segments may represent the stationary pieces of only noise or speech plus noise, and the length of each segment is variable and adapted automatically to the local stationarity of the data. Given the segmentation results, an EM procedure is then employed to update the maximum likelihood (ML) estimator of the noise and speech spectra as well as the speech waveform itself for each segment, using the estimates from previous segments as initial values. Because it relies on segments that represent the longest stationary pieces, this algorithm is able to maximize the statistical efficiency in producing accurate estimates while maintaining its adaptability to nonstationarity of the speech as well as the noise.

II. The SE Algorithm

Assume that the clean speech $x(t)$ is degraded by additive noise $v(t)$, so the corrupted signal $y(t)$ can be expressed as

$$y(t) = x(t) + v(t).$$ (1)

The objective of speech restoration is to estimate $x(t)$ from $y(t)$ when there is little or no prior knowledge of $v(t)$. To solve this problem, we propose an algorithm which consists of the following basic steps:

1. Obtain an optimal segmentation of $y(t)$ based on certain " stationarity" criteria, so that the segments represent the longest quasi-stationary pieces;

2. Process each segment sequentially as follows:

a) Refresh the noise spectral estimate if the segment does not contain speech;

b) Estimate $x(t)$ for each segment using an ML-EM procedure initialized by the spectrum of $y(t)$ in the segment and the spectral estimates obtained from previous segments;

3. Cascade the estimated speech segments to form the final estimate of the entire utterance $x(t)$.

The initial estimate of the noise spectrum can be obtained from the first segment under the assumption that it contains only the noise, as is typically the case.

Since the speech signal can be regarded as piecewise stationary, and in many cases, the statistical properties of environmental noise change slowly relative to the speech, the algorithm is able to make the maximum use of stationarity for accurate estimation and at the same time remain to be sufficiently flexible for tracking the time-varying speech and noise spectra.

We call this algorithm the *segmentation-estimation (SE) algorithm*. In the following, the segmentation and estimation steps are described in detail.

A. Segmentation

The aim of the segmentation step is to obtain the maximum-length quasi-stationary segments from the degraded speech signal. Typical SRS are based on fixed-length frames that are 10-20 ms long. Such short segments are necessary to capture transient sounds such as consonants. But they are typically too short for stationary sounds such as vowels. For accurate spectral estimation, it is necessary to combine as many short frames as possible in forming the longest stationary segments as the basic units of estimation.

Although ad hoc methods have been proposed for combing frames, we employ an entropy-based method proposed in [7-8]. In this method, the local cosine transform (LCT) is used to represent a nonstationary signal. The entropy of the LCT coefficients, as a complexity measure, can be used to represent the stationarity of the signal. Roughly speaking, a stationary signal can be represented by a simple model and therefore has a small entropy, whereas a nonstationary signal requires a more complex representation and hence a large entropy.

Therefore, an optimal segmentation of $y := \{y(t)\}$ can be obtained by minimizing the overall complexity of the LCT representation measured by

$$H(y) := \sum H(y_i) \tag{2}$$

where $\{y_1, y_2, \ldots\}$ form a partition of y.

To find the optimal segmentation, we employ the tree-based bottom-up procedure proposed in [7-8]. Given a signal y of length $N = 2^L$ (the zero-padding technique is applied if the length is not a power of 2), a

binary tree of depth $D \leq L$ is created by recursively bi-sectioning the time interval $[1, N]$. Each node of the binary tree represents a segment of y. For example, node (0,0) represents the entire signal, node (1,0) represents the left half of the signal, and node (1,1) represents the right half of the signal, etc. For each node, the LCT $\{c_i\}$ of the segment represented by the node is calculated; so is the entropy of the segment

$$H := -\sum p_i \log p_i \tag{3}$$

where $p_i := |c_i|^2 / \sum |s_j|^2$ and $\{s_j\}$ is the LCT of the entire signal y. The tree is then pruned from bottom up as follows: Two adjacent segments y_l and y_r of the same parent node will be combined to form a larger segment $y_p := \{y_l, y_r\}$ if and only if

$$H(y_l) + H(y_r) + \theta \geq H(y_p). \tag{4}$$

The cost of combining y_l and y_r is defined as

$$C(y_p) := H(y_l) + H(y_r) - H(y_p). \tag{5}$$

If it is decided to combine, the entropy of the resulting parent segment y_p is recomputed as

$$H(y_p) = H(y_l) + H(y_r) \tag{6}$$

and the pruning procedure continues until the largest segments reach a predetermined maximum length n_{max} which satisfies $2^D \leq n_{max} \leq N$. Note that $\theta > 0$ in (4) is a threshold which we introduce in this paper to take into account the statistical uncertainty of the entropy measurements. A large value of θ would result in more combinations of nodes and a small value of θ would lead to more small segments. So this parameter controls the degree of segmentation. The selection of this parameter will be discussed in the next section.

B. Estimation

With the nonstationary properly handled by the segmentation step, the estimation step computes the ML estimates of the speech and noise spectra sequentially for each of the quasi-stationary segments under the Gaussian assumption. It is a blind estimation method because it does not require prior knowledge of the speech or noise spectrum. The quasi-stationarity of the segments makes it possible to calculate the ML estimator by a frequency domain EM algorithm, where the "complete data" comprise $y(t)$ and $x(t)$. This algorithm is a generalization of the algorithm in [9] to include colored noise.

Let y be a segment of length n obtained from the segmentation step and let $Y(\tau)$ be the DFT of y. After k

iterations, let $S_x^{(k)}(\tau)$ and $S_v^{(k)}(\tau)$ be the estimated speech and noise spectrum, respectively. Then, the estimates at iteration $k+1$ are given by

$$S_x^{(k+1)}(\tau) := S_{x|y}^{(k+1)}(\tau) + n^{-1}|M_{x|y}^{(k+1)}(\tau)|^2 \qquad (7)$$

$$S_v^{(k+1)}(\tau) := n^{-1}|Y(\tau)|^2 \, S_{x|y}^{(k+1)}(\tau) / S_x^{(k+1)}(\tau) \qquad (8)$$

where

$$M_{x|y}^{(k+1)}(\tau) := \frac{S_v^{(k)}(\tau)Y(\tau)}{S_x^{(k)}(\tau) + S_v^{(k)}(\tau)} \qquad (9)$$

$$S_{x|y}^{(k+1)}(\tau) := \frac{S_v^{(k)}(\tau)S_x^{(k)}(\tau)}{S_x^{(k)}(\tau) + S_v^{(k)}(\tau)}. \qquad (10)$$

The estimate of x is given by the IDFT of $M_{x|y}^{(k+1)}(\tau)$. It approximates the minimum mean-square estimator $E(x|y)$. To initialize the iteration, one may use the spectral subtraction estimator

$$S_x^{(0)}(\tau) = \max\{\varepsilon, n^{-1}|Y(\tau)|^2 - \eta S_v^{(0)}(\tau)\} \qquad (11)$$

where $\varepsilon > 0$ is a small number and $a := \exp(-\eta/2)$ is interpretable as the false-alarm probability of speech detection at a given frequency. Moreover, $S_v^{(0)}(\tau)$ could be obtained from the most recent noise-only segment or an estimate from the previous segments.

III. IMPLEMENTATION

Now we discuss some practical issues concerning the implementation of the SE algorithm.

First, the threshold θ in (4) can be selected with the help of a graphical plot of the sorted costs in (5), an example of which is shown in Fig. 1. If the signal does comprise multiple homogenous segments, then it is expected that the cost function increases gradually at the beginning but starts to grow rapidly at a certain point. The cost at this point, which is approximately equal to $\exp(-5)$ in Fig. 1, can be used as θ. The threshold can be determined adaptively for each utterance or be fixed for all utterances. The latter may not result in optimal segmentation, but the computation complexity is lower. A fixed threshold is employed in our experiment.

One may take the periodogram $n^{-1}|Y(\tau)|^2$ as the initial guess $S_x^{(0)}(\tau)$ of the speech spectrum. But a better initial guess is given by (11), where the periodogram is soft-thresholded using the estimated noise spectrum to produce an improved (in the minimax sense) spectral estimator. Note that $\varepsilon > 0$ is small
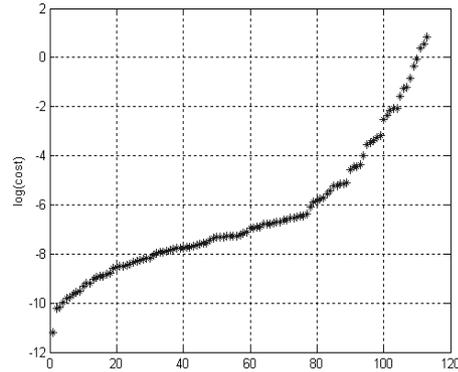

Fig. 1 Plot of sorted cost (in logarithm) for an utterance.

number employed to prevent the spectral estimate from being zero or negative and η is a tuning parameter.

For each segment, the initial guess $S_v^{(0)}(\tau)$ of the noise spectrum can be estimated from the most recent noise-only segment or can be taken as the estimate from the previous segment. In particular, it can be estimated from the first segment which is assumed to contain the noise only. This assumption is valid in most cases because the speakers don't speak immediately after the recognizer is turned on. Otherwise, a voice activity detector (VAD) can be used to find the first noise-only segment. In this case, all segments before the noise-only segment should be processed retrospectively. In our experiments we obtain $S_v^{(0)}(\tau)$ from the first segment.

IV. EXPERIMENTS

The SE algorithm is tested on a mixed database. Although the database consists of telephony data, the channel effect is not as important as the background noise, so equation (1) is a reasonable assumption.

The database contains 1800 utterances of Chinese names spoken by 120 speakers, each uttering 15 names. The speech signals are recorded in a quiet environment via telephone line, mobile or landline, sampled at 8k Hz. The database is both line-type and gender balanced. Two types of noise are considered. One is the bus noise recorded in a shuttle bus, which represents a typical noisy environment of mobile users in China. The other is the background noise recorded in a shopping mall. The noisy speech is formed by adding the noise to the clean speech at given SNR levels.

The speech recognition system we used features an effective search strategy and employs a rank-based continuous Gaussian HMM. It also employs a continuous spectral subtraction technique to handle possible noise corruption. It is based on a grammar (a set of rules), rather than a statistical language model, which contains about 1000 Chinese names and their baseform variations.

Table I contains the recognition rates in the bus noise environment at different SNR levels without speech enhancement. As can be seen, the sentence error rate (SER) increases nearly exponentially as the SNR decreases.

TABLE I
Recognition Rate Without Speech Enhancement

SNR (dB)	Sentence Error Rate (SER)
5	52 %
10	25 %
15	13 %
25	8 %

Tables II and III contain the results with speech enhancement using the proposed SE algorithm with different choice of $\eta = -2\log\alpha$. Four iterations are carried out on each segment. Except for the first segment which is assumed to contain only noise and used to obtained the initial guess for the noise spectrum, all other segments are treated equally as signal plus noise, even though some of them are actually noise-only segments. As shown in the tables, the relative improvement in the recognition accuracy is up to 28%, depending on the SNR and the choice of η. Note that the improvement shown in Table III is attributed solely to the EM iteration rather than the spectral subtraction in (11) because $\eta = 0$. Further improvement on these results is expected if a voice activity detector (VAD) is employed to distinguish the segments that contain speech from those that contain only noise where the noise spectrum can be updated.

TABLE II
Recognition Rate With Speech Enhancement: $\eta = .6$

SNR	SER	Relative Improvement
5	46 %	12 %
10	19 %	24 %
15	12.%	8 %
25	8 %	0 %

TABLE III
Recognition Rate With Speech Enhancement: $\eta = 0$

SNR	SER	Relative Improvement
5	47 %	10 %
10	18 %	28 %
15	12 %	8 %
25	7 %	13 %

TABLE IV
Recognition Rate Under Shopping Mall Noise

SNR	SER Without Speech Enhancement	SER With Speech Enhancement
5	99 %	90%
10	78 %	52%
15	31 %	24%

For further evaluation, Table IV shows the test results under the condition of shopping mall noise, which is the most damaging type of noise in all the scenarios we have investigated because of its close similarity with speech. As can be seen, improvement in the error rate by the proposed SE algorithm, which uses $\eta = 1$, is significant at all three SNR levels.

V. CONCLUDING REMARKS

We have proposed a segment-based algorithm, the SE algorithm, for the enhancement of noisy speech. Experiments have shown that the proposed algorithm is effective against the bus noise as well as the speech-like shopping mall noise.

We found that for sufficiently high SNR (e.g., ≥ 30 dB), employing the speech enhancement algorithm can hurt the performance. This, however, is not surprising because the statistical error in estimating negligible noise is known to be able to reduce or eliminate the benefit of estimation. The problem can be solved by switching off the speech enhancement algorithm when the SNR, estimated by an SNR estimator from the noisy signal, is sufficiently high.

As in [6] and [9], the estimation method in this paper can be generalized to include the estimation of an unknown channel. The effectiveness of this method is a subject for future research.

VI. References

[1] S. F. Boll, " Suppression of acoustic noise in speech using spectral subtraction," *.IEEE Trans. Acoust. Speech, Signal Processing,* vol. 27, pp. 113-120, Apr. 1979.

[2] J. S. Lim and A. V. Oppenheim, " Al-poll modeling of degraded speech," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. 26, pp. 197-210, June 1978.

[3] J. D. Gibson, B. Koo, and S. D. Gray, " Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Acoust. Speech, Signal Processing,* vol. 39, pp. 1732-1742, Aug. 1991.

[4] Y. Ephraim, " A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing,* vol. 40, pp. 725-735, Apr. 1992.

[5] K. Y. Lee and S. Jung, " Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Trans. Speech, Audio Processing,* vol. l8, pp. 282-291, Mar. 2000.

[6] Y. Zhao, " Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolusive noises," *IEEE Trans. Speech, Audio Processing,* vol. l8, pp. 255-266, Mar. 2000.

[7] R. R. Coifman and M. V. Wickerhauser, " Entropy-based algorithms for best basis selection," *IEEE. Trans. Inform. Theory,* vol. 38, pp. 713-718, Mar. 1992.

[8] E. Wesfreid and M. V. Wickerhauser, " Adapted local trigonometric transforms and speech processing," *IEEE Trans. Signal Processing,* vol. 41, pp. 3596-3600, Dec. 1993.

[9] A. K. Katsaggelos and K.-T. Lay, " Maximum likelihood identification and restoration of images using the expectation maximization algorithm," in *Digital Image Restoration*, A. K Katsaggelos ed., pp. 143-175, Springer-Verlag: Berlin, 1991.