Noisy Speech Recognition via Wavelet Transform Enhancement

Seokyong Moon, Jenq-Neng Hwang

Information Processing Laboratory
Dept. of Electrical Engr., FT-10
University of Washington
Seattle, WA 98195

Abstract

In the real world, the speech signals to be analysed are contaminated by background noise. Speech enhancement system is essential to make existing speech recogniser more robust to noisy environments. An enhancement system based on discrete wavelet transform (DWT) representation and solved by estimate-maximisation (EM) algorithm is proposed. This technique can find its counterpart in the Wiener filtering where Fourier transform or all pole modeling representation are employed.

1 Introduction

In the real world, noise sources are ubiquitous and we frequently encounter the situations in which the signals to be analysed are corrupted by background noise to some extent. The performance of existing speech recognition system severely degrades due to the presence of background noise in speech signal. Most speech recognition systems are designed based on the strong assumption about the environmental conditions in which the speech recognition systems are to be used. Most designs are predicated on assumptions that the environment is very quiet, the environmental conditions (e.g., noise characteristics) do not change over time. Therefore, a speech enhancement system is essential to enable the existing speech recogniser for practical use (see Figure 1).

It is our observation that the performance of a speech enhancement system is critically affected by the feature vectors that are selected among various representational schemes. Many speech enhancement systems use the linear predictive code (LPC) of speech signal because of its successful usage in automatic speech recognitions. Recently, representation based on discrete Wavelet transform (DWT) has led to the relaxation of the stationarity assumption of LPC representation, and is now available as a new alternative for speech enhancement [2, 7].

2 The Speech Recognizer

The speech recogniser (shown in Figure 1) used in this paper is a time delay neural network (TDNN) [8].

2.1 Time Delay Neural Network

A TDNN is a feed-forward type of network with a sigmoid function for each unit. Similar to the one used in [8], the TDNN employed in this paper has an input layer with 12 units, a hidden layer with 8 units, and an output layer with the same number of units corresponding to the same number of classes to be recognised. It remembers the previous outputs (e.g., maximum of 7 time-steps, or frames) of each unit. It is also fully interconnected from 5 frames of hidden layer with time delay 0, 1, 2 to the hidden units. It is also fully interconnected from 5 frames of hidden layer with time delay 0, 1, 2, 3, 4 to output units. There is no connection among the units of same layer and from the previous time-tap of higher layers (see Figure 2).

If we represent the weighted input to a unit by \( x_t \), the output of a unit \( y_t \), the index of a unit in the input, hidden, or output layer by \( i, j, \) or \( k \), respectively, then the equation for output \( y_k(t) \) of output unit \( k \) at time \( t \) can be represented by:
\[ y_h(t) = S(\sum_{j=1}^{12} \sum_{d=0}^{2} w_{jhd} S(\sum_{i=1}^{12} \sum_{m=0}^{2} w_{ijd} (t - l - d) + \theta_j) + \theta_h) \]  

where, \( d \) is the time delay from input to a hidden unit, \( l \) is the time delay from hidden to output units, \( w_{abc} \) is the weight from unit \( a \) to unit \( b \) along time-tap \( c \), \( \theta_a \) is the bias of unit \( a \), and \( S \) is a sigmoid function

\[ y = \frac{1}{1 + e^{-x}}, \quad S(.) = y_h(t), \quad S[.] = y_j(t - l). \]  

2.2 Time Delay Neural Network in Noise

Due to its prefixed amount of time delay span, the TDNN cannot easily recognize speech with significant amount of warping. Another drawback of a TDNN is its high sensitivity to ambient noise due to its deterministic nature of pattern matching. For example, the performance degrades very quickly if speech patterns are corrupted by noise.

We used in our simulations the speaker independent E-set data (B, D, G only), which is a subset of the ISOLET Spoken Letter Database collected by the Oregon Graduate Institute of Science & Technology [1]. The database consists of 7800 spoken letters, 2 productions of each letter by 150 speakers. To train the TDNN, 270 tokens of isolated words (B, D, G) from 45 female speakers are used. 180 tokens of isolated words (B, D, G) from another 30 female speakers are used for evaluating the performance. To test the performance in noisy environment, the white Gaussian noise of various dB levels is added to the testing E-set speech. The S/N is calculated from

\[ S/N(dB) = 10 \cdot \log_{10} \frac{\sum_{i=1}^{L} s(i)^2}{\sum_{i=1}^{L} d(i)^2} \]  

where \( L \) is the total length of speech data. The LPC coefficients of order 12 are used as features for each frames of speech pattern.

Figure 3(a) shows the performance of a TDNN for an isolated speaker independent word recognition task, where a TDNN is trained with noise-free speech and tested with noisy speech of various signal-to-noise ratio (S/N) levels. The performance degradation is clearly observed when the testing S/N drops below 40 dB which is nearly noise free condition. The 84% recognition accuracy at S/N of 50 dB drops to 33% accuracy at S/N of 0 dB. Figure 3(b) shows the performance of a TDNN trained with S/N of 5 dB noisy speech and tested with speeches of various S/N levels. It indicates that it can only provide 66% of accuracy in the vicinity of 5 dB S/N of speech (same level as the trained noisy speech) and it is worse at other S/N levels. The best performance attainable is an 18% decrease in accuracy when compared with the performance of the TDNN trained with noise-free speech.

Figure 2: The schematic architecture of a TDNN.

Figure 3: (a) The performance of a TDNN trained with noise-free speech and tested with noisy speech in various S/N levels. (b) The performance of a TDNN trained with S/N of 5 dB noisy speech and tested with speeches of various S/N levels.
3 Enhancement using Discrete Wavelet Transform

Wavelet theory [2, 7] emerges to be a unified framework for existing signal processing techniques, e.g., multiresolution signal processing, subband coding, and wavelet series expansion. The main interest of wavelet transform (WT) is the analysis of nonstationary signals. The WT does not predicate the assumption that the signal is stationary, or at least quasi-stationary, which provides a powerful tool for our temporal signal analysis, such as speech. It also provides a very efficient procedure, called discrete wavelet transform (DWT), for software implementation.

3.1 Orthonormal Discrete Wavelet Transform

An orthonormal wavelet transform of a signal $s(t)$ is defined by the synthesis/analysis equations:

$$z(t) = \sum_{m} \sum_{n} \bar{z}_m^n \cdot \psi_m^n(t) \quad (3)$$

$$\bar{z}_m^n = \int_{-\infty}^{\infty} z(t) \cdot \psi_m^n(t) dt \quad (4)$$

where $m$ and $n$ are the scale and shift indices, respectively. All orthonormal basis functions, $\psi_m^n(t)$, are dilation and translation of a single prototype function, $\psi(t)$:

$$\psi_m^n(t) = 2^{m/2} \cdot \psi(2^m t - n) \quad (5)$$

and the prototype wavelet $\psi(t)$ is commonly called a mother wavelet.

3.2 Enhancement using Wavelet Coefficient Thresholding

Motivated by the principal component analysis techniques for signal enhancement, where eigencomponent associated with small eigenvalues are regarded as noise and can be thresholded out, we that most DWT coefficients are small while important informations are kept in large values, we can devise a heuristic method based on thresholding the DWT coefficients. The estimate of DWT of signal, $s(t)$, is obtained by passing the DWT of noisy signal, $y(t)$, through the thresholding function $\eta_T(z)$.

$$\eta_T(z) = \begin{cases} z & \text{if } z > T \text{ or } z < -T \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

And then the estimate of signal $\hat{s}(t)$ is obtained by inverse DWT of $\bar{z}_m^n$. Figure 4 shows the rather poor performance of TDNN when DWT thresholding method is used as an enhancement system for speech recognizer. Although thresholding method is very simple and easy to implement, it is not an effective method for improving the performance of a speech recognition system. Furthermore, the performance critically depends on the threshold values which are difficult to determine.

3.3 Generalized Wiener Filtering of Discrete Wavelet Transform

A Wiener filter like estimation technique for a signal that is corrupted by additive white Gaussian noise (AWGN) is explored by adopting the framework of DWT. It takes a class of iterative procedure in signal estimation process. If the speech signal, $s(t)$, is corrupted by a zero-mean stationary AWGN, $d(t)$, which is independent of $s(t)$, then the observed signal, $y(t)$, can be represented by

$$y(t) = s(t) + d(t), \quad -\infty < t < \infty \quad (7)$$

and the problem of enhancement is to get the best estimate, $\hat{s}(t)$, of the original signal $s(t)$.

Due to the linear operations of DWT, the resulting observation wavelet transform coefficients have the variance which is defined by

$$y_m^n = \sigma_m^n + d_m^n \quad (8)$$

$$\text{Var}[y_m^n] = \sigma_m^2 + \sigma_d^2 \quad (9)$$

Our goal is to estimate the $s_m^n$ from $y_m^n$ by effectively exploiting the variances. Motivated by the classical Wiener filter theory, the minimum mean-square error (MSE) estimate of $s_m^n$ can be approximated by

$$\hat{s}_m^n = E[s_m^n|y_m^n] = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_d^2} \cdot y_m^n \quad (10)$$

and the optimal estimate of $s(t)$ can thus be obtained by inverse discrete wavelet transform (IDWT).
\[ \hat{s}(t) = \sum_{m,n} \psi^n_m(t) = \sum_{m,n \in R} \left[ \frac{\sigma^2_m}{\sigma^2_m + \sigma^2_n} \right] y^n_m(t) \] (11)

To have a good estimate of \( \sigma^2_m \), according to Equation 10, a good estimate of \( \sigma^2_m \) is desired first. The latter can be formulated as a statistical maximum likelihood (ML) problem \[4\], and solved by the iterative optimization procedure suggested by Lim & Oppenheim \[5\], where an iterative Wiener filtering based on LPC spectrum (all pole modeling) was proposed to solve the speech enhancement problem. This iterative optimization procedure can find its basis in the statistical estimate maximization (EM) algorithm introduced by Dempster et. al. \[3, 4\]. Our proposed DWT method can be vaguely denoted as a generalised Wiener filtering in DWT domain, where the DWT coefficient at each scale is treated as a frequency response. The iterative procedure can be summarised as:

1. Discrete wavelet transform of noisy speech, \( \{y^n_m\} \).
2. Estimate the variances, \( \{\sigma^2_m\} \) and \( \{\sigma^2_n\} \).
3. Design a generalised Wiener filter.
4. Perform the generalised Wiener filtering on DWT coefficients.
5. Go back to Step 2, until converge.
6. Reconstruct the speech via inverse DWT.

3.4 Simulation Results
Very encouraging enhancement results based on the proposed DTW generalised Wiener filtering were obtained. For an objective comparison of the proposed speech enhancement technique, the DWT enhanced speech is recognised by the TDNN after LPC preprocessing. Figure 5 shows the favorable performance of the DNN recognition results of the proposed DTW generalised Wiener filtering approach when compared with the conventional all pole modeling \[5\] and spectral subtraction techniques \[6\].

To have a simple illustration of the DWT enhanced speech, Figure 6 shows the enhancement result of an isolated noisy speech of 5dB S/N. Also as shown in Figure 7, a continuous noisy speech of -5dB S/N can also be enhanced with acceptable quality. Furthermore, the proposed technique seems to improve intelligibility when compared with the all pole modeling technique \[5\] in a subjective human listening test.

References
Figure 6: (a) Noise free isolated utterance "B" of a female speaker. (b) Noisy speech with 5dB S/N. (c) Enhanced speech.

Figure 7: (a) Noise free continuous speech of "Robert hasn't bought the car, but he will later today". (b) Noisy continuous speech of -5dB S/N. (c) Enhanced speech.