UTILIZING INTERBAND ACOUSTICAL INFORMATION FOR MODELING STATIONARY TIME-FREQUENCY REGIONS OF NOISY SPEECH

Chang D. Yoo

Korea Telecom
17 Woomyeon-dong, Secho-gu, Seoul, 137-792

ABSTRACT

A novel enhancement system is developed that exploits the properties of stationary regions localized in both time and frequency. This system selects stationary time-frequency (TF) regions and adaptively enhances each region according to its local signal-to-noise ratio (LSNR) while utilizing both the acoustical knowledge of speech and the masking properties of the human auditory system. Each region is enhanced for maximum noise reduction while minimizing distortion. This paper evaluates the proposed system through informal listening tests and some objective measures.

1. INTRODUCTION

In speech enhancement, the main objective is to maximally reduce noise while minimizing speech distortion. To attain such objective, *a balanced tradeoff between noise reduction and speech distortion must be achieved, as noise reduction invariably introduces speech distortion.* Many enhancement methods in the past sought this tradeoff over the entire spectrum of short segments of fixed length. However, by exploiting both time- and frequency-localized behavior of speech and also utilizing both the acoustical knowledge of speech [1, 2, 3, 4] and the masking properties of human auditory system, a tradeoff that better achieves its objective is possible.

In this paper, a system which exploits the properties of stationary regions localized in TF is described. As in [5], the proposed system identifies and selects stationary TF regions using M-band decomposition with adaptive analysis windowing. For each selected region, the system makes various parameter adjustments for maximal noise reduction with minimal distortion. Instead of sequentially processing the channels as in [5] without utilizing any interband acoustical knowledge, the proposed system enhances the M channels in parallel using interband acoustical information.

The two essential operations involved in the tradeoff mentioned above are spectrum estimation and noise reduction. When estimating the speech spectrum, an all-pole model is often used. Given that the resolution of an all-pole spectrum is determined by the order of the model [6], the order can be adjusted to achieve an appropriate spectral resolution to suite the local condition of a region. Many enhancement methods have relied on fixed order model regardless of both the signal characteristics and the SNR of each speech segment when estimating the speech spectrum; however, by varying the model order according to the changing characteristics of speech a more suitable estimate can be made. When reducing noise, the Wiener filter has often been used although with limited success. By modifying the Wiener filter and adjusting its parameters to reflect both the varying behavior of speech and LSNR, a more balanced tradeoff is achieved.

As described in [5], the system makes no effort to decimate and then interpolate the subbands to reduce computation as in many of the systems described in [7] since the number of subbands is limited to only a few- a band per kilo-Hertz for maximum performance- the computational saving by decimation is only marginal (factor of number of subbands) and the difficulty of eliminating or compensating for aliasing can be overbearing.

The paper is organized as follows: Section 2 describes the proposed overall enhancement system. Sections 3 describes how interband acoustical information is used to detect unvoiced TF regions. Section 4 presents some examples using the proposed system. Finally, Section 5 concludes the paper.

2. ENHANCEMENT SYSTEM

This system enhances speech in three steps. The first is identifying and selecting stationary TF regions in degraded speech by M-band decomposition with adaptive analysis windowing. The second is estimating the spectrum and then adaptively enhancing each selected region according to its local signal-to-noise ratio (LSNR) while utilizing the acoustical knowledge of speech. A modified Wiener filter based on selective linear prediction (SLP) model is used to enhance each region. By adjusting both the order of the model and the parameters of the filter to suit the local characteristics, each region is enhanced for maximum noise reduction and minimum distortion. The third step involves synthesizing the M enhanced channels and ultimately the enhanced signal.

2.1. Overview

The overall system is shown in Figure 1. As shown, degraded speech $y[n]$ is initially decomposed into M channels $\{y^{(k)}[n]\}_{k=1}^{M}$ by an M-band filter bank $\{H^{(k)}(\omega)\}_{k=1}^{M}$ such that $y^{(k)}[n] = \sum_i h^{(k)}[n-i]y[i]$ where $h^{(k)}[n]$ is the impulse response of the k^{th} channel filter with frequency response $H^{(k)}(\omega)$. The passband and non-passband of $H^{(k)}(\omega)$ are denoted respectively by $R_H^{(k)}$ and $R_{H-}^{(k)}$ so $R_H^{(k)} \cup R_{H-}^{(k)} = \Omega_s = [0 \quad \pi]$. Henceforward, the superscript k will refer to the k^{th} channel. For example, $y^{(1)}[n]$ refers to the dc channel, and $y^{(k-1)}[n]$ and $y^{(k)}[n]$ are adjoining channels for $k = 2, \ldots, M$. The decomposition satisfies the following three

0-7803-5041-3/99 $10.00 © 1999 IEEE

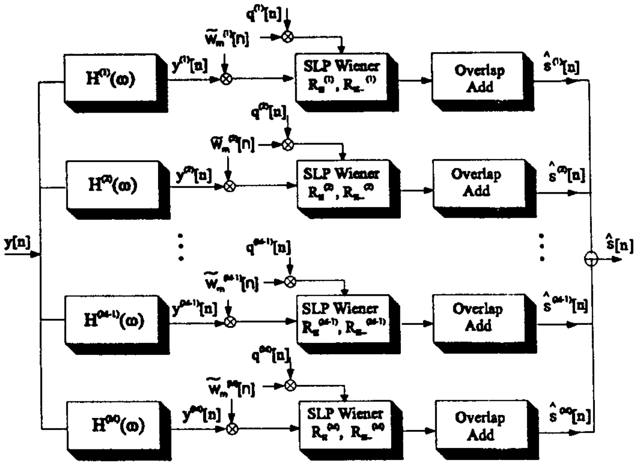

Figure 1: Overall enhancement system.

conditions [1]:

$$\sum_{k=1}^{M} H^{(k)}(\omega) = e^{j\omega n_o} , \tag{1a}$$

$$R_H^{(k)} \cap R_H^{(k+1)} \approx \emptyset , \quad k = 1, \ldots , M-1 \tag{1b}$$

$$\bigcup_{i=1}^{M} R_H^{(i)} = \Omega_s . \tag{1c}$$

Once the decomposition satisfying the above conditions is performed, stationary regions in each channel are identified and selected using an adaptive window whose length varies according to the changing spectral characteristics of the channel[2]. The adaptive window $\tilde{w}_m^{(k)}[n]$ corresponding to the m^{th} time interval and k^{th} channel is designed such that the following condition is satisfied:

$$\sum_m \tilde{w}_m^{(k)}[n] = 1 , \qquad \forall n . \tag{2}$$

The normalized cross-correlation between the smoothed spectra in two different time intervals is used as a similarity measure; thus, signal within an analysis interval will have steady similarity measure [8].

Following the selection, each region is enhanced using the modified Wiener filter based on an all-pole spectrum. To estimate the all-pole spectrum in the frequency range of interest - either in passband or non-passband regions- selective linear prediction (SLP) is used. The spectral resolution is adjusted by varying the model order to suite both the LSNR and the acoustical behavior of the region. In addition to varying the order of the model, the parameters of the modified Wiener filter are also adjusted to suite local conditions. In order to determine the local conditions, the k^{th} reference signal $q^{(k)}[n] = y^{(1)}[n]$ is used (discussed in Section 3). The enhancement is performed on a frame-by-frame basis where a frame constitutes a windowed segment of varying length. The m^{th} frame of the k^{th} channel denoted by $y_m^{(k)}[n]$ is given by

$$y_m^{(k)}[n] = y^{(k)}[n] \cdot \tilde{w}_m^{(k)}[n] . \tag{3}$$

The modified Wiener filter of the k^{th} channel and m^{th} time interval is denoted by $\Gamma_W^{(k)}(m,\omega)$ and is given by

$$\Gamma_W^{(k)}(m,\omega) = \frac{P_s^{(k)}(m,\omega)}{P_s^{(k)}(m,\omega) + c^{(k)}(m) \cdot P_z^{(k)}(m,\omega)} ,$$

where $P_s^{(k)}(m,\omega)$ and $P_z^{(k)}(m,\omega)$ are respectively the estimated SLP spectrum of speech and the noise spectrum. To estimate $P_s^{(k)}(m,\omega)$, separate sets of SLP coefficients are used for regions $R_H^{(k)}$ and $R_{H-}^{(k)}$. As mentioned above, the order and the number of iteration are dependent on both the acoustical nature and the LSNR. The parameter $c^{(k)}(m)$ is also varied depending on the local conditions.

The overlap-add method is used for the synthesis of each channel so that $\hat{s}^{(k)}[n]$ (see Figure 1) is the enhanced signal of $y^{(k)}[n]$ for $k = 1, \ldots , M$, and once each channel is synthesized, $\{\hat{s}^{(k)}[n]\}_{m=1}^{M}$ are summed for the synthesis of the enhanced signal $\hat{s}[n]$.

3. THE USE OF ACOUSTICAL INFORMATION OF SPEECH

In order to improve the intelligibility of the overall speech, the enhancement of both voiced and unvoiced sounds have to be performed equally well. Unfortunately, the energy of unvoiced sound is much lower than that of voiced, and often in noisy speech unvoiced sound is inaudible in noise while voiced sounds are perfectly audible. This is reason why it is so difficult to enhance unvoiced sounds. By incorporating some acoustical knowledge of speech, the proposed system can improve the enhancement of unvoiced sounds.

Considered to be mid-to-high frequency noise, unvoiced sound is characterized by the spectral location of the energy weight. The

[1] n_o: constant

[2] A stationary signal is used here to roughly mean a signal whose frequency content does not vary with time.

Table 1: Summary of modeling the region of i^{th} channel and m^{th} time interval of length N. **v**:voiced region and **uv**:unvoiced region

LSNR$^{(i)}(m)$	order	$c^{(i)}(m)$	Num. iter.
> 15dB	N/3	0.5	3
0dB-15dB	v: N/3 uv: 2	v: 0.75 uv: 0.5	2
< 0dB	0	v:1.5 else 3.0	3

proposed enhancement system determines whether a signal localized in TF region is unvoiced or not by comparing its power to that corresponding to the baseband: the k^{th} channel of m^{th} time interval $y_m^{(k)}[n]$ where $k > 1$ is considered *unvoiced* when its power $P^{(k,k)}(m)$ satisfies the following condition[3]:

$$P^{(k,k)}(m) > Q \cdot P^{(1,k)}(m)$$

where

$$\begin{aligned} P^{(i,j)}(m) \\ &= \frac{1}{\Delta_H^{(i)}}\Big(\sum_n |y^{(i)}[n]\tilde{w}_m^{(j)}[n]|^2 \\ &\quad - \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} S_{zz}^{(i)}(\Omega)|W_m^{(j)}(\Omega-\omega)|^2 d\omega d\Omega\Big). \end{aligned} \tag{4}$$

where $S_{zz}^{(k)}(\omega)$ and $W_m^{(k)}(\omega)$ respectively represent k^{th} channel noise spectrum and Fourier transform of $\tilde{w}_m^{(k)}[n]$. When noise exhibits similar acoustical behavior as unvoiced, the enhancement of unvoiced sound becomes a matter of identifying and preserving the spectral location of the energy weight by reducing less noise around the spectral location of the energy weight relative to other spectral regions. In order to determine if a region is unvoiced or not, $q^{(k)}[n] = y^{(1)}[n], k = 1, \dots, M$ is required in order to evaluate (4). see Figure 1).

4. EXAMPLE

To illustrate the performance of the proposed system two clean sentences - "That shirt seems too long" and "He has the bluest eyes"- respectively spoken by a female speaker and a male speaker, sampled at 10kHz are degraded at an SNR of 10dB by additive white Gaussian noise and are then enhanced using parameters shown in Table 1.

In Figures 2 and 4, each shows spectrograms of clean, noisy and enhanced. Figures 3 and 5 show plots of segmental SNR of both enhanced and noisy. The figure shows that noise reduction not only in silence regions but also in regions where various acoustical forms of speech are present- of course speech distortion is kept to a minimum for regions where speech is present. The segmental SNR is defined as follows:

$$\text{seg. SNR of } \hat{s}[n] \text{ at } m = 10\log_{10}\left(\frac{\sum_{n=0}^{N_s} s^2[n-m]}{\sum_{n=0}^{N_s}(\hat{s}[n-m]-s[n-m])^2}\right). \tag{5}$$

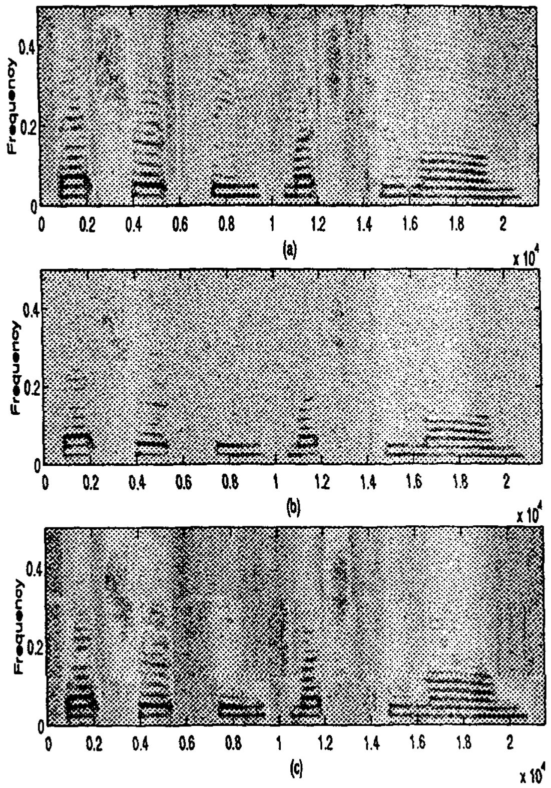

Figure 2: Spectrograms of (a) clean ("That shirt seems too long") (b) noisy and (c) enhanced.

5. CONCLUSIONS

The problem of reducing noise in speech which has been degraded with additive noise has been investigated. The purpose of this study was to develop a system that would maximize noise reduction while minimizing speech distortion. To attain this goal, a balanced tradeoff between noise reduction and speech distortion must be target-ted, as noise reduction often leads to speech distortion. Traditional enhancement methods try to achieve this balance over the entire spectrum of a fixed-length windowed speech segment; however, by exploiting the local characteristics of stationary TF regions and utilizing both interband acoustical information of speech and the masking properties of the human auditory system, a tradeoff that better achieves its objective can be made.

[3] $\Delta_H^{(k)} = \frac{\int_{-\pi}^{\pi}|H^{(k)}(\omega)|d\omega}{\sum_i \int_{-\pi}^{\pi}|H^{(i)}(\omega)|d\omega}$.

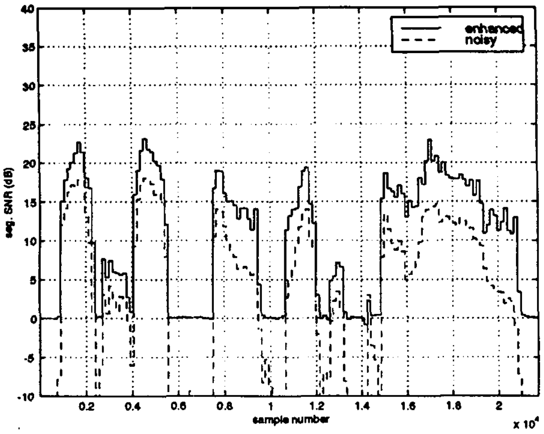

Figure 3: Segmental SNRs of enhanced and noisy (shown in Figure 2) versus sample number using Equation 5 with $N_s = 150$.

6. REFERENCES

[1] J. John R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, p. 27.5.1. New York, New York: McMillan, 1993.

[2] D. O'Shaughnessy, *Speech Communication.* Massachusetts and New York: Addison Wesley., 1987.

[3] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* New Jersey: Prentice-Hall., 1978.

[4] B. S. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. ASSP-23, pp. 201-212, June 1976.

[5] C. D. Yoo, "Selective all-pole modeling of degraded speech using m-band decomposition," *ICASSP*, May 1996.

[6] J. Markel and J. A.H. Gray, *Linear Prediction of Speech.* Berlin: Springer-Verlag., 1976.

[7] P. Vaidyanathan, *Multirate Systems and Filter Banks.* New Jersey: Prentice Hall Signal Processing Series, 1993.

[8] C. D. Yoo, *Speech Enhancement: Identification and Modeling of Stationary Time-Frequency Regions.* PhD thesis, MIT, E.E.C.S. Department, August 1996.

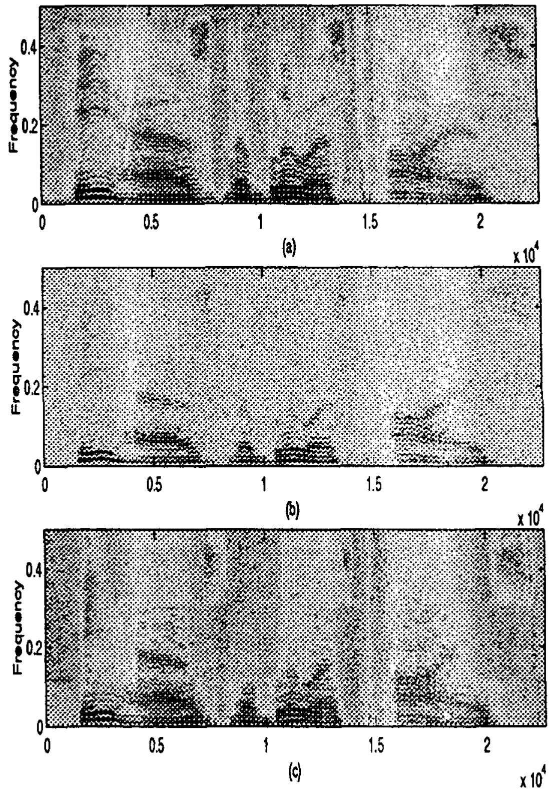

Figure 4: Spectrograms of (a) clean ("He has the bluest eyes")(b) noisy and (c) enhanced.

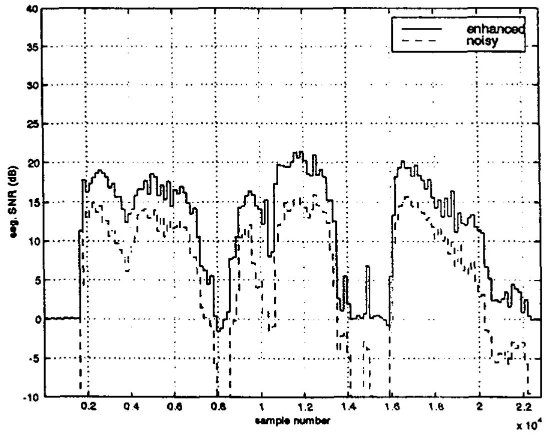

Figure 5: Segmental SNRs of enhanced and noisy (shown in Figure 4) versus sample number using Equation 5 w ith $N_s = 150$.