Proc. IEEE Int. Symp. Circuits, Systems (Singapore), pp. 312-315, June 1991

ADAPTIVE POSTFILTERING FOR ENHANCEMENT OF NOISY SPEECH IN THE FREQUENCY DOMAIN

Peter Kabal[1,2], Fang-Ming Wang[2], Douglas O'Shaughnessy[2] and Ravi P. Ramachandran[1]

[1]Electrical Engineering
McGill University
Montreal, Quebec H3A 2A7

[2]INRS-Télécommunications
Université du Québec
Verdun, Quebec H3E 1H6

Abstract

This paper presents a new frequency-domain adaptive postfilter for enhancement of noisy speech. The postfilter suppresses the noise in spectral valleys and allows more noise in the formant regions where it is masked by the speech signal. First, we perform an LPC analysis of the noisy speech and calculate its log magnitude spectrum. After identifying the formants and spectral valleys, the log magnitude spectrum is modified to obtain the postfilter frequency response. This response has local minima in the regions corresponding to the spectral valleys and local maxima of equal magnitude at the formant frequencies. The filtering uses an overlap-add FFT strategy. Experimental results show that this new frequency-domain approach results in enhanced speech of better perceptual quality than obtained by a time-domain method. Our method is especially efficient in eliminating high frequency noise and in preserving the weaker, high frequency formants in sonorant sounds.

1. Introduction

The quality and intelligibility of speech is often degraded by background acoustic noise, coding noise, distortion due to transmission errors, or interference from background speakers. The aim of speech enhancement is to process the degraded speech such that its quality and intelligibility are improved. Our approach is to use an adaptive postfilter.

Consider a typical spectrum of a speech signal that has both formant peaks and spectral valleys. For speech degraded by additive white noise, it is known that the noise in the frequency regions corresponding to the valleys contributes the most to perceptual distortion. The role of a postfilter is to (1) accurately track the time-varying nature of speech and (2) suppress the noise in the spectral valleys. The frequency response of a postfilter corresponds to a modified version of the speech spectrum in which (1) there are local minima or dips in the regions corresponding to the spectral valleys and (2) local maxima of equal magnitude at the formant frequencies. The dips will suppress the noise, thereby accomplishing noise reduction. Using spectral peaks of equal magnitude at the formant frequencies ensures that there is no additional lowpass tilt in the output signal (after postfiltering), and allows for relatively more noise in the formant regions. Some speech distortion is introduced because the signal levels in the formant regions are altered due to the postfiltering. There is a tradeoff between noise reduction and speech distortion [1]. The filter is adaptive due to the time-varying nature of speech.

The approach in [1] can be classified as a time-domain method in that the postfiltering is implemented as a time difference equation. The frequency response of the postfilter approximates a modified version of the spectrum of the noiseless input speech. In this paper, we develop a frequency-domain approach to accomplish postfiltering in which the postfilter is represented by a set of DFT coefficients. The motivations for using a frequency-domain approach are: (1) the frequency-domain method allows for independent control over different portions of the spectrum, especially those corresponding to the formant locations and spectral valleys, (2) by observing the time evolution of the spectrum, the method allows for the suppression of regions corresponding to low energy or silence. Our experimental results show that our new method performs better than the method described in [1].

2. Time-domain Methods

In the time-domain approach, the transfer function of the postfilter $H(z)$ is based on an LPC model,

$$H(z) = \frac{A_M(z/\beta)}{A_N(z/\alpha)} \text{ where } A_P(z) = 1 - \sum_{i=1}^{P} a_i z^{-i}, \qquad (1)$$

i.e., $A_P(z)$ is the inverse filter of a Pth order autoregressive process. The postfilter $H(z)$ accomplishes noise reduction by suppressing the noise around the spectral valleys, but distorts the speech signal by sharpening the formant peaks. There is no lowpass filtering effect. The postfilter is adaptive in that the LPC coefficients a_i are updated in each frame; either forward or backward adaptation can be used. In [1], fixed values of α and β are used, $0.5 \le \alpha \le 1$ and $\beta < \alpha$.

3. Frequency-domain Postfiltering

Fig. 1 shows a block diagram of frequency-domain postfiltering. The postfilter is represented by its DFT coefficients $H(k)$, which are multiplied by $P(k)$, which is a modified form of $X(k)$ (the DFT coefficients of the input noisy speech $x(n)$). The filtering of the input speech is performed in the frequency domain. An inverse DFT yields the postfiltered signal $y(n)$. The DFT is sufficiently long to allow $y(n)$ to represent a linear convolution of $p(n)$ and $h(n)$. The input speech, sampled at 8 kHz, is divided into frames of 128 samples (16 ms).

3.1 Calculation of the Log Magnitude Spectrum

An approximation of the speech spectrum is obtained by calculating the log magnitude spectrum of $1/A_P(z)$. First, we determine the LPC coefficients a_i and hence the filter $A_P(z)$. In each frame of speech, we use the autocorrelation method with a Hamming window of length 256 samples to perform a 16th order analysis. This frame length includes 2–3 pitch periods in order to obtain accurate spectral estimates [2]. A 16th order analysis allows us to resolve 3–4 formants. Experiments have shown that a higher order analysis results in too many peaks in the spectrum, thereby making it difficult to identify the formant locations.

Given $A_P(z)$, the first step is to obtain $A_P(k)$ $(k = 0, \ldots, 255)$ which is a N_{PT}-point FFT of the sequence $\{1, -a_1, -a_2, \cdots, -a_p\}$. The log magnitude spectrum is $R(k) = -20 \log_{10} |A_P(k)|$. This is used for identifying the formants.

3.2 Formant and Valley Identification

The log magnitude LPC spectrum $R(k)$ (see Fig. 1) is computed for a frame of speech corrupted by noise. Finding the amplitude and location of the formants is an important step in determining the postfilter coefficients $H(k)$. Formant extraction is simpler for clean (noiseless) speech [3][4] than for our noisy speech (SNR typically $\le$ 10 dB). For the majority of speech segments that have most of

CH 3006-4/91/0000 – 0312 $1.00 © IEEE

their energy at low frequencies, noise dominates at high frequencies and the upper formants may go undetected.

We use a peak picking strategy to detect the formants. Given $R(k)$, we sequentially determine the local maxima and decide whether to classify a given peak as a formant. In each frame of speech, we find a maximum of four formants. Two major problems with peak picking are that [4]: (1) some peaks may be spurious and (2) two formants may appear as one peak. Deciding whether a peak corresponds to a formant is necessary to avoid classifying spurious peaks as formants. The second problem of merged peaks is not crucial for implementing a postfilter since a formant region containing merged peaks is sharpened anyway by the postfilter.

We first determine a maximum energy level ($A_{\max}$) and a noise level (N_{av}); $A_{\max} = \max(R(k))$, for $k = 0, \ldots, N_{PT}/2$. The location (or the value of k) at which $A_{\max}$ occurs is denoted by $L_{\max}$. L_{av} approximates the average noise level. Consider frames with noise only. For each of these frames, we calculate $A(m)$ as the sum of the amplitudes of the peaks of $R(k)$ divided by the number of peaks. Then, N_{av} is the average of ten values of $A(m)$.

3.2.1 Detection of Unvoiced Segments

In analyzing each frame of the noisy speech: (1) each frame is classified as unvoiced ($L_u = 1$) or not-unvoiced ($L_u = 0$), (the category not-unvoiced includes voiced speech and pure noise), (2) the formant amplitudes and locations are determined, (3) the amplitudes and locations of the spectral valleys are found.

To identify the unvoiced frames of speech, we do the following: (1) set $L_u = 1$, (2) if $A_{\max} < 2N_{\text{av}}$, then $L_u = 0$; stop, (3) if $L_{\max} < N_{PT}/4$, then $L_u = 0$; stop, (4) calculate A_{MX} (see below), (5) if $A_{MX} < N_{\text{av}}/2$, then $L_u = 0$; stop. (The thresholds for comparison in Steps (2) and (5) were chosen after examining many frames of speech.)

The criterion in Step (2) indicates a frame of either pure noise or weak speech (a segment with low energy). For voiced speech (either strong or weak), the largest peak which is a formant occurs below 2 kHz. In unvoiced speech, most of the energy is at high frequencies and the largest peak occurs between 2.5 and 4 kHz [3]. Satisfying the criterion in Step (3) indicates voiced speech. There are rare cases when pure noise having much energy at high frequencies (like noise bursts) will not satisfy the criteria in Steps (2) and (3). To discriminate between this case and a truly unvoiced segment, a quantity A_{MX} is calculated by partitioning the frequency range into four equal regions comprising 32 samples each. Then, $A_{MX} = \max(S_{R1}, S_{R2}, S_{R3}, S_{R4})$, where the S_{Rn} are average values of $R(k)$ in each of the regions. The major difference between unvoiced segments and pure noise is that peaks in unvoiced segments have a much wider bandwidth than those in pure noise. Noise having a spectrum with a narrow peak at a high frequency is undetected by the criteria in Steps (2) and (3). Therefore, the value of A_{MX} is higher for unvoiced segments.

3.2.2 Formant Amplitudes and Locations

Since our speech is bandlimited to 3.4 kHz, we only examine $R(k)$ in its first 110 points to detect the formants. The strategy is to sequentially examine each value of $R(k)$, locate a peak, and decide if it is a formant. Formant detection is invoked when a local peak in $R(k)$ is found, under the constraint that $R(k) > R_{\min}$. Two possibilities emerge: (1) if more than one formant has been found so far, Part 1 is invoked, (2) if at most one formant has been detected, only Part 2 of the algorithm is invoked. In each frame of speech, the algorithm results in a total of N_F formants being found. The amplitudes of these formants are placed in the array $A_P(J)$. The index locations of $R(k)$ at which these formants occur are stored in the array $N_P(J)$.

Part 1 of the Formant Detection Algorithm

1. If $A_{\max} \geq C_2 N_{\text{av}}$ and $A_{\max}/A_P(2) \leq C_1$, a formant is found. Otherwise, proceed to Part 2.

Part 2 of the Formant Detection Algorithm

1. If both parts below are satisfied, a formant is detected. Otherwise, go to Step (2).
   $$A_{\max} > C_3 N_{\text{av}} \text{ and } A_{\max}/R(k) \leq C_5$$
2. If $A_{\max} \leq C_3 N_{\text{av}}$, then a formant is found if both parts below are satisfied. Otherwise, go to Step (3).
   $$A_{\max} > C_1 N_{\text{av}} \text{ and } A_{\max}/R(k) \leq C_4$$
3. If $A_{\max} \leq C_3 N_{\text{av}}$ and $A_{\max} \leq C_1 N_{\text{av}}$, then a formant is found if $L_u = 1$ (unvoiced frame). Otherwise, the local peak in $R(k)$ does not correspond to a formant.

Discussion of the Algorithm

Most of the steps in the formant detection algorithm involve comparisons of $A_{\max}$ and N_{av} to differentiate among strong speech, medium level speech, weak speech, unvoiced speech, and pure noise before classifying a peak as a formant. Both $A_{\max}$ and N_{av} depend on the SNR in that, as the SNR increases, $A_{\max}$ increases and/or N_{av} decreases. The algorithm uses thresholds in each of the steps. The thresholds C_1, C_2 and C_3 are empirically chosen depending on the SNR. As SNR increases, C_1, C_2 and C_3 diminish. The values of the thresholds C_4 and C_5 are primarily chosen to detect the second formant and are sometimes used to establish higher formants.

For Part 1 of the algorithm (two or more formants already located), in a frame of voiced speech, we may encounter the problem in which the higher order formant peaks are hidden by the noise component. The tests attempt to pick up these high frequency formants. We accept a candidate peak as a formant automatically if the peak energy is high enough compared to the noise, but not too high compared to the second formant peak. If there is a big difference between the F1 and F2 peaks, higher-frequency formants are likely to be unreliably extracted in the noise background. If the two tests in Part 1 are negative (i.e., weak peak energy, or a big difference between F1 and F2 peaks), the candidate peak must pass through the tests of Part 2.

The satisfaction of the first condition in Part 2, Step (1) indicates a frame of strong speech (peak energy $A_{\max}$ well above the noise level N_{av}). Then, the second half of Step (1) indicates the presence or absence of a formant (the formant candidate must be sufficiently high compared to the highest peak); this primarily detects the second formant (i.e., $A_{\max}$ refers to the first formant peak). Part 2, Step (2) indicates neither strong nor weak speech (i.e., average speech). The second part of the test determines whether the peak corresponds to a formant. Note that the threshold in Step (2) is smaller than that of Step (1). Step (3) tests for weak speech, unvoiced speech, or pure noise. For Step (3), the peak in $R(k)$ corresponds to a formant only if the frame was identified as unvoiced.

Formant Adjustment

The algorithm above finds N_F formants in each frame. We introduce an additional modification that adjusts the number of formants to ensure continuity in the formant trajectories from frame to frame. We encountered instances where the number of formants detected in a particular frame is suddenly less than the number in those previous frames. This problem is dealt with as follows.

Suppose, in a particular frame m, $N_F = 1$. If $N_F \geq 3$ in frame $m-1$ and $N_F \geq 2$ in frame $m-2$, there is an abrupt discontinuity in the formant trajectories. If the value of $A_{\max}$ in frame m is approximately that in frame $m-1$, the value of N_F is adjusted to be 2. The location of the second formant $N_P(2)$ in frame m is set to the location of the third formant peak in frame $m-1$. The corresponding amplitude $A_P(2) = R(N_P(2))$ for $R(k)$ calculated in frame m. A formant peak is reinserted at a relatively high frequency so that the postfiltering operation results in the recovery of the high frequency components of the speech signal. Similarly, an adjustment in the value of N_F from 2 to 3 in frame m is made if (1) $N_F \geq 4$ in frame $m-1$, (2) $N_F \geq 3$ in frame $m-2$ and (3) the values of A_{MAX} in frames m and $m-1$ are approximately equal. Now, $N_P(3)$ is set to the fourth formant in the previous frame. Then, $A_P(3) = R(N_P(3))$. Lastly, the formant detection algorithm can occasionally pick a total of five formant peaks; then, the peak with the lowest amplitude is discarded. The method of formant adjustment in a particular frame is based on the number of formants detected in the previous two frames. Since there is no provision for looking ahead at succeeding frames, this approach is suitable for real-time processing.

3.3 Modification of the Log Magnitude Spectrum

The log magnitude spectrum $R(k)$ is modified to become $S(k)$ such that in the postfiltered speech, the formant peaks are sharpened, the spectral valleys are deepened, and no unwanted lowpass tilt is present. We first divide $R(k)$ into sections from $k = 0$ to $N_P(1)$, $N_P(1)$ to $N_P(2)$, $\cdots$, $N_P(N_F)$ to 109, and finally 110 to 128. Each section is individually modified. This freedom of independently modifying different sections of $R(k)$ is the advantage of the new frequency-domain approach over the time-domain method.

The postfilter coefficients $H(k)$ must be determined from the modified log magnitude spectrum $S(k)$. Note that $S(k) = 20\log_{10}|H(k)|$. The phase of $H(k)$ is the same as the phase of $1/A_P(k)$. The postfilter coefficients are obtained by modifying only the magnitude of the LPC spectrum.

3.4 Modification of $X(k)$ - Smooth Switching Algorithm

In many speech utterances, there are transitions of very weak speech or silence to or from frames having a relatively stronger speech component. For speech degraded by noise, these transitions contain a substantial noise component. Usually, no formants are detected in these transition regions, and postfiltering is not useful. We use another strategy: we use a smooth switching algorithm to attentuate $X(k)$ in a region of very weak speech or pure noise ($N_F = 0$) that arises between frames having a stronger speech component ($N_F > 0$).

The frames for which $N_F = 0$ are classified into one of three possible states. If no formants are found in the current frame m and $N_F > 0$ in frame $m-1$, frame m is said to be in the state (0,1), which indicates a transition from speech to pure noise. The next 7 frames are also assigned to state (0,1) if no formants are detected there. In any section of the utterance for which $N_F = 0$, there is a maximum of 8 frames in state (0,1).

After 8 frames in state (0,1), additional frames in which $N_F = 0$ do not correspond to a transition from speech to pure noise, but are indeed segments of very weak speech or pure noise. These frames are assigned to state (0,0), which is further subdivided. State (0,0,0) corresponds to pure noise in which there is no indication of a transition to frames with a speech component. State (0,0,1) generally corresponds to very weak speech and provides an indication of a transition to strong speech. We face the problem of knowing in advance when $N_F > 0$ will be encountered. Before speech frames appear, there are usually some frames of very weak speech at low frequencies in which no formants are detected. Such frames are in state (0,0,1). To distinguish between states (0,0,0) and (0,0,1), consider $R(k)$ for $k = 5, \ldots, 20$ (i.e., 150 to 500 Hz). If $R(k) > 1.2N_{\text{av}}$ for any k between 5 and 20, the corresponding frame is assigned to state (0,0,1). Otherwise, the frame is assigned state (0,0,0).

Finally, consider the frames in which formants are detected ($N_F > 0$), i.e., either in state (1,0) or (1,1). Suppose $N_F > 0$ in frame m and $N_F = 0$ in frame $m-1$. Then, frame m is in state (1,0), a state in which there is a transition from a segment of very weak speech or pure noise to a segment with a stronger speech component. All other frames for which $N_F > 0$ are in state (1,1).

The method for obtaining $P(k)$ from $X(k)$ depends on the state of each frame. First, consider the frames in state (0,1). Given M consecutive frames numbered $L = 1, 2, \cdots, M$ in state (0,1), we define a modification factor $D = 0.8 - 0.1L$. Then,

$$P(k) = \begin{cases} D\,X(k) & k = 5 \text{ to } 16 \\ 0 & \text{otherwise} \;. \end{cases} \tag{2}$$

The low frequency components are gradually deemphasized. For frames in state (0,0,0), $P(k) = 0$ for all k. For state (0,0,1), there is an indication that a strong speech component will appear soon. However, modifying $X(k)$ to allow for a smooth transition to frames with a strong speech component is difficult since there is no *a priori* knowledge as to when these frames will appear. Since state (0,0,1) usually corresponds to weak speech at low frequencies, we introduce the following scheme to preserve some low frequency components:

$$P(k) = \begin{cases} 0.3X(k) & k = 5 \text{ to } 16 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

If a frame is either in state (1,0) or (1,1),

$$P(k) = \begin{cases} 0 & k = 0, 1 \text{ or } k = 110, \ldots, 128 \\ X(k) & \text{otherwise.} \end{cases} \tag{4}$$

Noise at frequencies beyond 3400 Hz and at very low frequencies is eliminated.

3.5 Generation of the Postfiltered Output Signal

Postfiltering is performed only for frames with a relatively strong speech component (states (1,0) and (1,1)). For frames in state (1,0), the signal $y(n)$ is scaled by 0.2. Experiments have shown that this abruptness is noticeable if the scaling is not done. For states (0,0,0), (0,0,1) and (0,1), no postfiltering is done, i.e., $Y(k) = P(k)$. The final postfiltered output $v(n)$ is obtained by introducing a 50% overlap between the 256-sample segments of $y(n)$ and adding the corresponding samples (an overlap-add strategy).

4. Experimental Results

Fig. 2 shows wideband spectrograms of noisy and enhanced speech for the sentence 'Cats and dogs each hate the other' for a male speaker. Fig. 2(a) shows the effect of added noise at 10 dB SNR. The strong first formant is visible throughout for the vowels, but the weaker second formant disappears in the word 'each.' Higher-frequency formants are only visible for strong vowels with a high first formant (e.g., F3 and F4 in 'cats,' 'dogs'). Frication (e.g., the 'ts' in 'cats') is totally obscured by the noise. Fig. 2(b) shows the results after enhancement in the time domain; where the speech is strong enough (i.e., during vowels), noise is suppressed in the non-formant regions, primarily at high frequencies. However, no effect is seen during the non-vowel portions of the speech, and the background noise is as strong as ever there. In Fig. 2(c), we see the effect of our frequency-domain enhancement; the noise is significantly suppressed during the non-vowel portions of the signal. In addition, some frication is identified properly and retained in the output (e.g., the aspiration of /k/ in 'cats,' and the final frication in 'each'). Furthermore, the higher formants are in general better modeled in the frequency-domain approach than in the time-domain method; e.g., F3 and F4 in 'dogs,' F2–F4 in 'hate,' and F2 in 'the other.' Experiments with the same sentence spoken by a female reveal the same observations as for a male speaker. Informal listening tests clearly indicate a preference for the frequency-domain method over the time-domain method. The much decreased noise level with the frequency-domain approach leads to much more pleasant speech, while retaining as much as possible of the phonetic information to keep intelligibility high.

5. Summary and Conclusions

This work has formulated a new frequency-domain approach for adaptive postfiltering of noisy speech. In this approach, a new method to detect the formants and spectral valleys of the speech spectrum is introduced. Based on the locations of the formants and spectral valleys, the DFT coefficients of the postfilter are determined for the purposes of suppressing the noise around the spectral valleys and sharpening the formant peaks. Experimental results show that the perceptual speech quality is improved with the new method compared to the time-domain postfiltering method.

References

1. V. Ramamoorthy, N.S. Jayant, R.V. Cox and M.M. Sondhi, "Enhancement of ADPCM speech coding with backward-adaptive algorithms for postfiltering and noise feedback", *IEEE J. on Select. Areas Commun.*, vol. SAC-6, pp. 364–382, Feb. 1988.
2. D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, 1987.
3. J. D. Markel, "Digital inverse filtering - A new tool for formant trajectory estimation", *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 129–137, June 1972.
4. S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 135–141, April 1974.

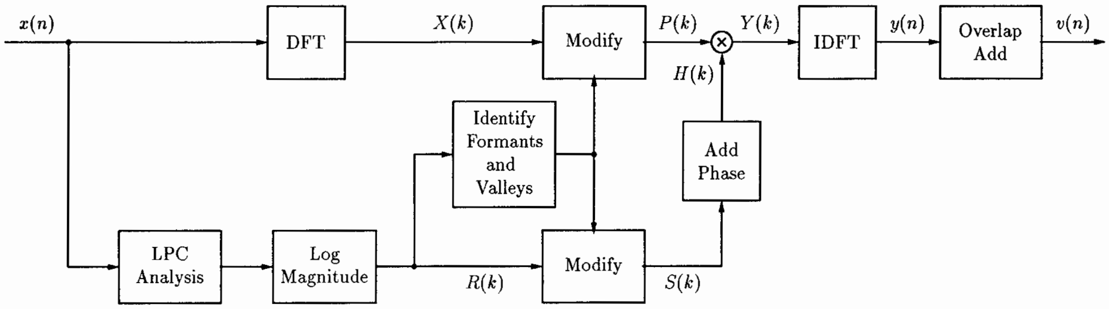
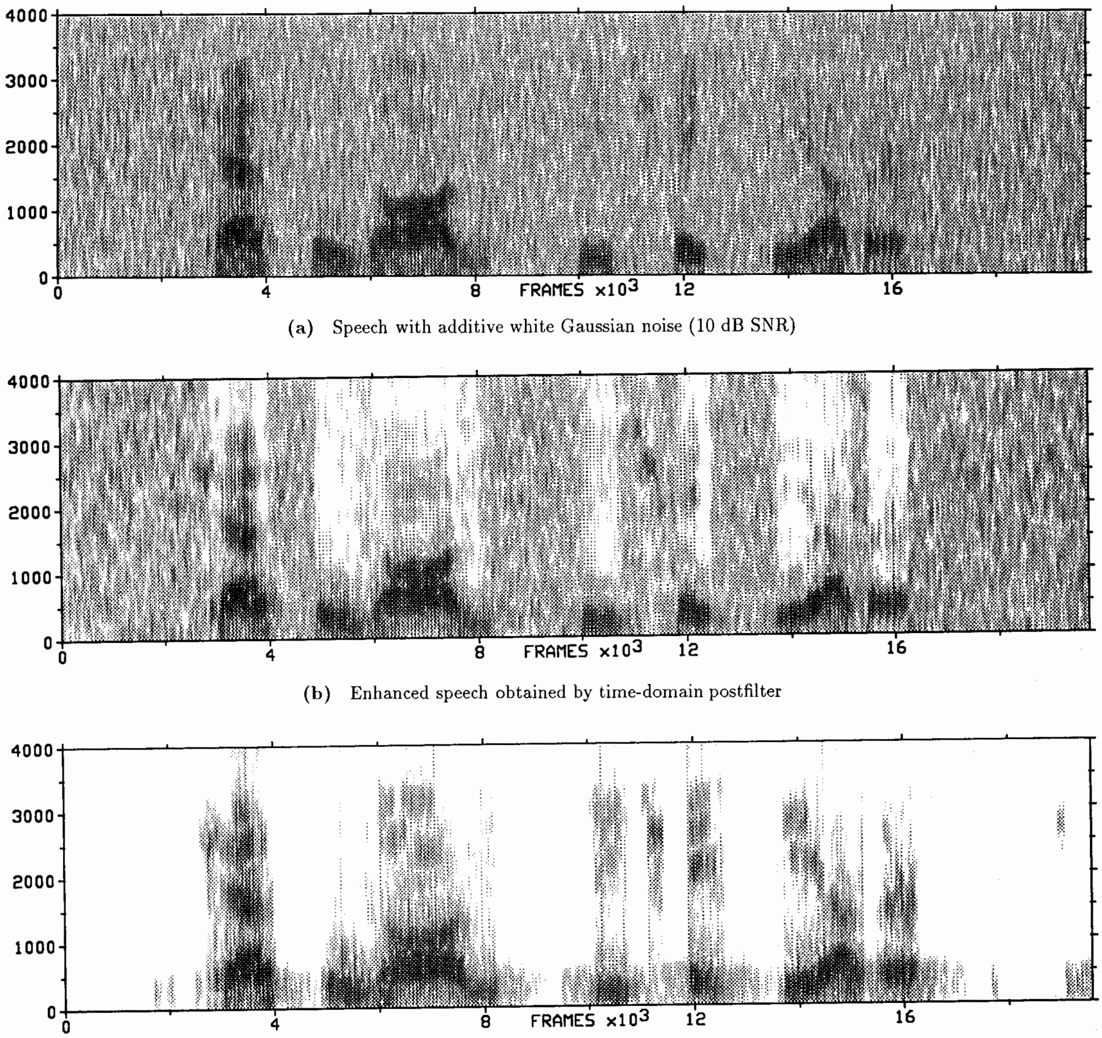


Fig. 1 Postfiltering in the frequency-domain



(a) Speech with additive white Gaussian noise (10 dB SNR)

(b) Enhanced speech obtained by time-domain postfilter

(c) Enhanced speech obtained by frequency-domain postfilter

Fig. 2 Wideband spectrograms of noisy and enhanced speech