... ENHANCEMENT BY SPECTRAL SUBTRACTION

Felix ALBU, Neculai DUMITRIU, Livia Doina STANCIU

University "POLITEHNICA" Bucharest

Faculty of Electronics and Telecommunications

B-dul Iuliu Maniu nr 1-3, sector 6 , Bucuresti

Tel : 410 54 00

... Spectral subtraction is performed over frames by obtaining the short-term magnitude ... noisy speech spectrum, subtracting an estimated noise magnitude spectrum from the ... spectrum, and inverse transforming this spectral amplitude using the phase of the ... speech. The technics presented in this paper differ by the window used for frame ... the power exponent, the weighted subtraction coefficient. The results have been ... compared by informal listening and SNR computations.

Introduction

In practice, recorded or transmitted speech contain an amount of noise. The ... posses a disproportionate amount of linguistic information. Because vowels ... posses larger amounts of energy, broadband noise degradation tends to mask ... sections more than voiced, thus causing decreased intelligibility. This ... coupled background noise has a spectral density depending on the ... conditions during recording . However, in most cases the additive noise ... simulated as white. The background noise causes a degradation of speech which ... to total unintelligibility. Two approaches dominate in the literature: the first is ... on the use of spectral subtraction, and the second on appropriate filters for noise ... [2].

Spectral subtraction is a *family* of frequency-domain noise reduction techniques ... on direct estimation of the short-term spectral magnitude. In this approach, speech ... modeled as a random process to which uncorrelated random noise is added. It is ... that the noise is short term stationary, with second-order statistics estimated ... silent frames. The estimated noise power spectrum is subtracted from the ... transformed noisy input signal. A generalized estimator is given by:

$$\hat{S}_s(\omega,m) = \left[\left|S_y(\omega,m)\right|^a - k\cdot\left|\hat{S}_d(\omega,m)\right|^a\right]^{1/a}\cdot e^{j\varphi_y(\omega,m)} \qquad (1)$$

where $|S_y(\omega,m)|$ and $\varphi_y(\omega,m)$ are obtained from the short-term Discrete Time Fourier Transform (stDTFT) of the noisy speech frame; $|\hat{S}_d(\omega,m)|$ is the short-term magnitude spectrum of the noise that must be updated during the absence of the speech; $\hat{S}_s(\omega,m)$ is the estimated stDTFT of the frame of speech from the spectral [illegible] the power exponent and k is the weighted subtraction coefficient

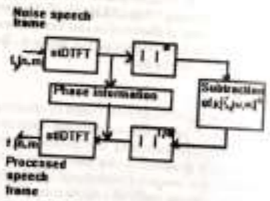


Fig. 1. Generalized spectral subtraction

The resultant spectrum is converted to the speech signal by using the [illegible] original degraded speech.

The speech waveform was sampled at a rate of 8000 Hz. The 180 [illegible] current frame were overlapped with the 76 trailing samples of the [illegible] through trapezoidal windowing (Fig. 2.)

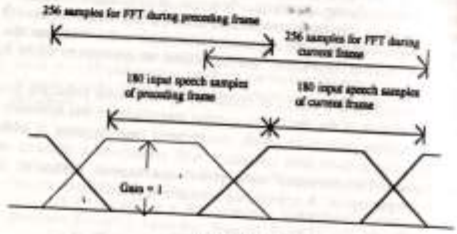


Fig. 2. Frame overlapping

We chose a frame of 76 samples because the resulting 256 samples permit [illegible] use of a standard FFT for the time-to-frequency transformation. The speech samples [illegible] overlapped so that the sum of amplitude weights of the overlapped windows is unity [illegible] indicated in Fig. 2 , a portion of the speech samples is used twice in the time [illegible] frequency transform: the trailing end of the preceding FFT and the leading edge of [illegible] current FFT frame. Then, a two-way transforms involving FFT and inverse FFT [illegible]

[illegible] audio effects are generated such as: clicking, [illegible] the speech signal. We propose another window (composed [illegible] is given by:

$$w(n) = \begin{cases} \frac{1}{1+e^{\text{[illegible]}}} & 1 \le n \le L \\ 1 & L+1 \le n \le N-L \\ \frac{1}{1+e^{\text{[illegible]}}} & N-L+1 \le n \le N \end{cases} \quad (2)$$

[illegible] the length of the window, p is the slope of the sigmoidal function, and L is [illegible] normalized magnitude in dB for the composed sigmoidal windows (p = [illegible] in Fig. 3b. For any window there are two desirable features:

[illegible] mainlobe width;

[illegible] attenuation for the sidelobes;

[illegible] window has a relatively narrow mainlobe width (Fig. 3b) which decreases [illegible] depends on the slope p and the overlap L, and the sidelobe attenuation is not [illegible]

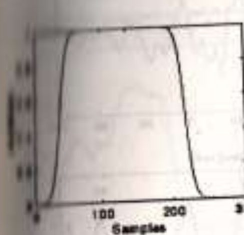


Fig. 3a. Composed sigmoidal window (N=256, L=76, p =0.2)

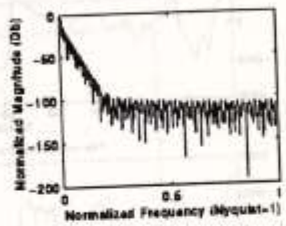


Fig. 3b. Normalized magnitude in dB for the composed sigmoidal windows(N=256, L=76, p=0.2)

From (1) it can be noticed that the estimated speech magnitude spectrum is not [illegible] guaranteed to be positive. Different systems remedy this by performing half-wave [illegible] rectification or full-wave rectification). We used half-wave rectification (i. e. negative [illegible] portions are set to zero). Forcing negative spectral magnitude values to zero, however, introduce a "musical" tone artifact in the reconstructed speech ... the major limitation of spectral subtraction technique ... improvement at low SNR (Fig. 5) but introduced a "musical" ... reconstructed speech.

Both windows (trapezoidal window and composed sigmoidal ...) comparable performance in removing of undesirable audio effects ... terms of SNR for presented windows shows comparable results with ... for voiced frames in case of composed sigmoidal window with low ...

Our results showed that if the weighted subtraction coefficient ... musical tone artifacts can be reduced. It is desirable to adjust k to ... and maximum spectral floor based on the estimated input SNR. The optimum ... must be between 1 and 3, with lower value for augmented segmental SNR ... the performance of the system decreases.

The results showed that spectral subtraction might increase the ... processed signal for a large range of the parameter a. The processed speech ... equal to 1 or 0.5 sounded "less noisy" at relatively high SNR. The results ... show an improvement of SNR especially for unvoiced frames.

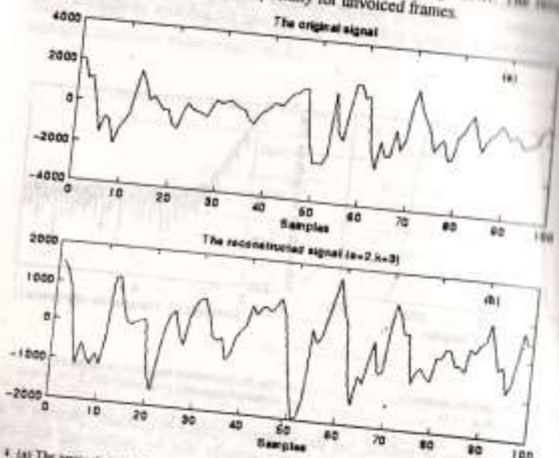
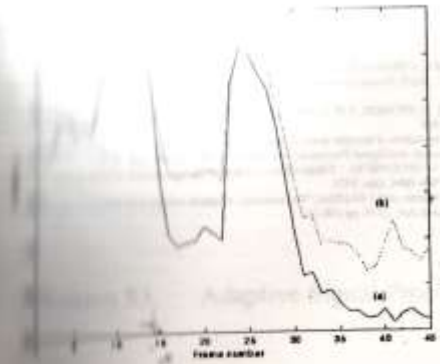


Fig. 4. (a) The original signal for a frame of the phoneme "A") (b) The reconstructed signal (a=2, k=3, p=0.1, global SNR= 0 dB)



... of the noisy speech (a), and the reconstructed speech (b) for the frames of the Romanian word ... (a=2, k=1, p=0.2, global SNR = 0 dB)

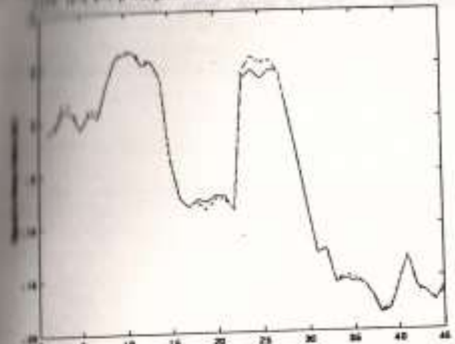


Fig. 6. The segmental SNR for the various frames of a processed Romanian word "SASE" with trapezoidal window (solid line) and composed sigmoidal window (dashed line) (k=1, p=0.02, Global SNR = 0 dB)

REFERENCES

[1]. G. S. KANG and L. J. FRANSEN, "Quality Improvement of LPC-Processed Noisy Speech
Subtraction " , *IEEE Transactions on Acoustics, Speech, and Signal Processing*,
June , 1989

[2]. J. R. DELLER, J. G. PROAKIS, J. H. L. HANSEN, "*Discret-Time Processing of Speech*
New York, 1993

[3]. S. F. BOLL, "Suppression of acoustic noise in speech using spectral subtraction ", *IEEE*
Acoustics, Speech, and Signal Processing, vol. ASSP-27, PP. 113-120, Apr. 1979

[4] J. S. LIM and A. V. OPPENHEIM, " Enhancement and bandwidth compression of noisy speech ",
vol. 67, pp. 1586-1604, Dec. 1979

[5].M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise ",
ICASSP .Record, Apr. 1979, pp 208-211.