ISCA Archive
http://www.isca-speech.org/archive

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

Speech Enhancement via Energy Separation

Hesham Tolba and Douglas O'Shaughnessy
Institut National de la Recherche Scientifique, INRS-Télécommunications, Québec, Canada.
E-mail: tolba@inrs-telecom.uquebec.ca and dougo@inrs-telecom.uquebec.ca.

Abstract

This work presents a novel technique to enhance speech signals in the presence of interfering noise. In this paper, the amplitude and frequency (AM-FM) modulation model [7] and a multi-band analysis scheme [5] are applied to extract the speech signal parameters. The enhancement process is performed using a time-warping function $\beta(n)$ that is used to warp the speech signal. $\beta(n)$ is extracted from the speech signal using the Smoothed Energy Operator Separation Algorithm **(SEOSA)** [4]. This warping is capable of increasing the SNR of the high frequency harmonics of a voiced signal by forcing the the quasiperiodic nature of the voiced component to be more periodic, and consequently is useful for extracting more robust parameters of the signal in the presence of noise.

1 Introduction

The problem of enhancing speech degraded by additive background noise has received considerable attention in the past two decades. The success of an enhancement algorithm depends on the goals and assumptions used in deriving the approach. The objective of achieving higher quality, more intelligibility, or reducing listener fatigue depends on the specific application. The principal application areas of speech enhancement are noise reduction for human listening, preprocessing for recognition systems, and preprocessing for linear predictive coding.

In this paper, we introduce a novel speech enhancement technique based on the AM-FM modulation model [4]. This enhancement technique is based on the fact that the speech signal can be modeled as the sum of N AM-FM signals. This model represents each component of the speech signal as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure. In our speech enhancement system, this model is used to facilitate both the analysis of the noisy speech signal and the extraction of the speech parameters that can help in calculating the time-warping function that will be used to warp our original speech signal in order to enhance it.

Our goal in this paper is to enhance the speech signal by enhancing its voiced component via time-warping. This could be achieved by applying a time-warping function, extracted from the original speech signal, to its voiced component. The effect of the warping function is to force the quasiperiodic nature of the voiced component to be more periodic, which increases the SNR of the voiced component [8] and consequently enhances the speech signal. One we obtained the enhanced modified voiced speech signal, all speech applications that use the **SEOSA** [2]-[7] are used to extract more robust parameters from the speech signal. Moreover, the voiced part could be enhanced using a comb filtering, while the unvoiced part could be enhanced using spectral subtraction techniques in order to enhance the overall speech signal.

The outline of this paper is as follows. In section (2) we describe and review the AM-FM Modulation Model, the energy operator, and the SEOSA. Then in section (3), an iterative algorithm to calculate the time-warping function is described, and the system, where both sinusoidal synthesis and interpolation are employed to enhance the speech signal, is presented. Experimental results that demonstrate the effectiveness of our algorithm are presented in section (4). Finally, in section (5) we conclude and discuss our present and future work.

2 AM-FM Modulation Model

Modeled as a sum of AM-FM signals, the speech signal can be processed easily to estimate several parameters such as the amplitude of the envelope, the instantaneous frequency of each resonance (peak) at each time instant t, the tracking of the formants and pitch extraction. Towards the extraction of these parameters, isolation of individual resonances (peaks) by bandpass filtering the speech signal around its resonances (peaks) must be performed. Then these parameters can be estimated for each resonance (peak) using an energy tracking operator such as **SEOSA**.

The model described in this paper represents each component of a speech signal as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure. That is, a speech signal

$s(t)$ can be defined as:

$$s(t) = \sum_{i=1}^{N} a_i(t)cos[\underbrace{2\pi f_i(t+\beta(t))+\theta}_{\phi_i(t)}], \qquad (1)$$

where N represents the number of peaks of the speech signal spectrum, $a_i(t)$ and $\phi_i(t)$ are the amplitude and frequency modulation functions of the i^{th} AM-FM signal component and $\beta(t)$ represents the frequency modulating signal. The instantaneous frequency $f_{inst}(t)$ of each component of the speech signal is defined as

$$f_{inst_i}(t) = \frac{d}{dt}[\phi_i(t)] = 2\pi f_i(1+\frac{d}{dt}[\beta_i(t)]) \qquad (2)$$

The discrete-time speech signal using the AM-FM signal model can be represented by:

$$s(n) = \sum_{i=1}^{N} a_i(n)cos[2\pi f_i(nT+\beta_i(nT))+\theta], \qquad (3)$$

where $a_i(n)$ and $\phi_i(n)$ are the discrete amplitude and frequency modulation functions of the i^{th} AM-FM signal component. The phase $\phi_i(n)$ is defined as

$$\phi_i(n) = \overline{\omega}_i[n+\beta(n)]+\theta_i = \overline{\omega}_i n^{'} + \theta_i \qquad \forall i, \quad (4)$$

where $\overline{\omega}_i$ is the average peak frequency for the window time interval and θ_i is the phase offset. The time-warping function $\beta(n)$ is such that the mapping $n^{'} = g(n) = n+\beta(n)$ is a nonlinear warping of the measured quasiperiodic signal into a periodic signal in $n^{'}$ with period $2\pi/\overline{\omega}_i$.

2.1 Smoothed Energy Operator Separation Algorithm (SEOSA)

This energy operator first developed by Kaiser [1] is defined as

$$\Psi[x(n)] \triangleq x^2(n) - x(n-1)x(n+1), n = 0, .., N-1 \quad (5)$$

In [2] it was shown that, when the energy operator Ψ is applied to an AM-FM signal

$$x(t) = a(t)cos[\int_0^t \omega_i(\tau)d\ \tau], \qquad (6)$$

it can approximately estimate the squared product of the amplitude and frequency signals; i.e.,

$$\Psi[x(t)] \approx [a(t)\omega_i(t)]^2 \qquad (7)$$

assuming that $a(t)$ and $\omega_i(t)$ do not vary too fast with time compared to the carrier frequency ω_c.
Demodulation of each AM-FM component of the speech signal into its amplitude envelope $|a_i(t)|$ and instantaneous frequency signal $\omega_i(t)$ is performed using Energy separation as follows [3], [4]:

$$\frac{\Psi[x(n)]}{\sqrt{\Psi[x\cdot(n)]}} \approx |a_i(n)|, \qquad \sqrt{\frac{\Psi[x\cdot(n)]}{\Psi[x(n)]}} \approx \Omega_i(n), \quad (8)$$

where $\Psi[x(n)]$ is the Teager-Kaiser energy operator and dots denote time derivatives and $\Omega_i = \omega_i T$.
After analysis in [3], [4], it was found that for AM-FM signals, Equation (8) could be written as

$$\begin{aligned} \sqrt{\frac{\Psi[x(n)]}{1-(1-\frac{\Psi[x(n)-x(n-1)]}{2\Psi[x(n)]})^2}} &\approx |a_i(n)|, \\ cos^{-1}(1-\frac{\Psi[x(n)-x(n-1)]}{2\Psi[x(n)]}) &\approx \Omega_i(n). \end{aligned} \quad (9)$$

The precise result from applying the energy operator to an AM-FM signal is given in [3],[4]. The energy approximation error can be reduced without affecting the desired envelope and frequency by filtering the energy operator $\Psi[x(n)]$ through an appropriate LPF. Using a seven-point binomial smoothing filter with impulse response $(1, 6, 15, 20, 15, 6, 1)$ the energy approximation error decreases typically by 50% [6], [7]. Consequently the envelope and frequency estimation errors are reduced.

2.2 Multi-Band AM-FM Demodulation

By the selection of an appropriate bandpass filter, isolation of an AM-FM signal component could be possible [5]. This is due to the fact that the wide-band FM signal could be limited to a limited-band signal without loosing the original signal if the limited-band signal contains more than 98% of the total power of the original FM signal. Moreover, noise components not falling within the vicinity of the desired local AM-FM component could be rejected.
To isolate the local modulation energy of an AM-FM signal component, it is necessary to utilize a bank of bandpass filters centered at each peak of the speech signal with an appropriate bandwidth. A neighboring spectral peak that has not been eliminated through bandpass filtering can seriously affect the estimated envelope and instantaneous contours. Resonance isolation is performed using Gabor filters having an impulse response $g(t) = exp(-\alpha^2 t^2)\cos(\omega_c t)$, where $\omega_c = 2\pi f_c$ is the resonance frequency and α is the parameter's bandwidth. α must be chosen as wide as is possible to include the FM signal around the peak, but narrow enough to exclude those of neighboring FM signal components. Bandpass filtering is implemented using a truncated, discretized Gabor filter [5].

3 Multi-band SEOSA Speech Enhancement

The speech is first windowed using Hamming window of 32 ms, with 16 ms displacement. The windowed speech is transformed into the frequency domain using an N-point FFT. Then the Energy operator is calculated for each peak in the Fourier spectrum by isolating each peak using Gabor bandpass filters centered around each peak. Finally, the instantaneous frequency $\omega_i(n)$ for each peak is estimated via the SEOSA. The instantaneous phase $\phi_i(n)$ is the integral of $\omega_i(n)$. Then the time-warping function $\beta_i(n)$ can be found by solving:

$$\phi_i(n) = \overline{\varpi}_i[n+\beta(n)] + \theta_i = \overline{\varpi}_i n^{'} + \theta_i \qquad \forall i. \qquad (10)$$

After warping, the speech signal can be modeled as a time-warped signal as follows:

$$s(n) = \sum_{i=0}^{N} a_i(n) cos[\overline{\varpi}_i(n+\beta(n)) + \theta]; \qquad (11)$$

where $\overline{\varpi}_i$ is the average instantaneous frequency for the i^{th} resonance and θ_i is its phase offset. The time-warping function $\beta(n)$ is such that the mapping $n^{'} = g(n) = n + \beta(n)$ is a nonlinear warping of the measured quasiperiodic signal into a periodic signal in $n^{'}$ with period $2\pi/\overline{\varpi}_i$, where $\overline{\varpi}_i$ is the average fundamental frequency over a window duration, and given by:

$$\overline{\varpi}_i = \frac{1}{|t_b - t_a|} \int_{t_a}^{t_b} \omega_i(t) d(t). \qquad (12)$$

An iterative technique for estimating $\beta(n)$ based on a more simplified version of the algorithm described in [9] has been developed which serves to get an accurate estimate of $\beta(n)$. The more accurate the estimate of $\beta(n)$, the more that we can improve and enhance the speech signal by warping it using $\beta(n)$. Once the initial average fundamental frequency is estimated using Equation (12), $\beta(n)$ can be found by solving:

$$\phi_i(n) = \overline{\varpi}_i[n+\beta(n)] + \theta_i = \overline{\varpi}_i n^{'} + \theta_i \qquad \forall i, \qquad (13)$$

Then $\beta_i(n)$ is used to warp the original speech signal to obtain the signal modeled using Equation (11).

Two possible strategies can be used to get the enhanced speech signal. The first constructs the enhanced speech signal via synthesis from the parameters already calculated. The resultant speech signal $\hat{s}(n)$ is synthesized on a frame by frame basis, as shown in Figure (1), according to:

$$\hat{s}_j(n) = \sum_{i=0}^{N} a_{i,j}(n) cos[\overline{\varpi}_{i,j}(n+\beta_j(n)) + \theta_{i,j}]; \qquad (14)$$

where j is the frame index and the constants $\theta_{i,j}$ are adjusted such that the synthesized signal is smooth across frames, i.e., the phase and frequency for each frame are adjusted to smoothly interpolate between the parameters which are estimated for successive segments. The second strategy, which was adopted in our analysis, is performed via an interpolation processes in discrete time according to the relation $n^{'} = n + \beta(n)$ by interpolating new signal samples at the new uniform discrete-time sample points $n^{'}$ to get:

$$\hat{s}_j(n^{'}) = \sum_{i=0}^{N} a_{i,j}(n^{'}) cos[\overline{\varpi}_{i,j} n^{'} + \theta_{i,j}]. \qquad (15)$$

Once we get $\hat{s}_j(n)$, a new iteration can be performed on the synthesized signal to enhance the $\beta_j(n)$ estimate until convergence, i.e, the speech signal is the most periodic. The effect of this warping increases the SNR of the voiced component and consequently enhances the speech signal.

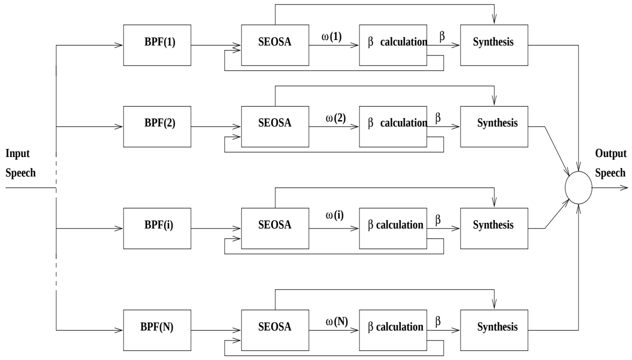

Figure 1: Multi-band AM-FM Speech Enhancement System

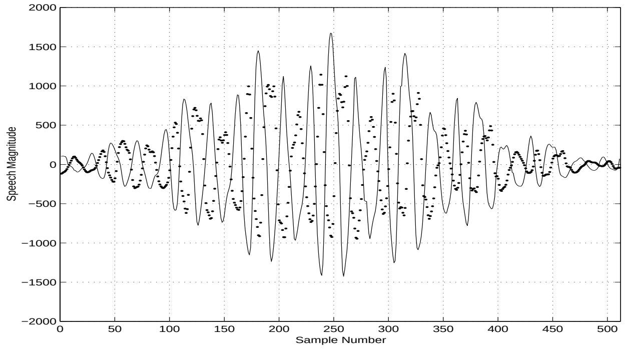

Figure 2: The time domain waveform of a speech sentence, represented by *dot* symbols, from the NTIMIT database before warping versus the waveform of the same sentence after warping.

4 Experimental Results

The speech corpus for this experiment is a subset of the NTIMIT database. The speech was sampled at 16 kHz. The NTIMIT database is the telephone bandwidth version of the widely used TIMIT database. Several kind of noise can be found in the NTIMIT database, such as broadband noise, band-limiting, low frequency hum, crosstalk, dial pulses, shot noise and sharp pulses.

Our algorithm can be summarized as follows. Each frame of 512 samples (32 msec) is weighted by a 512-point Hamming window, and then the DFT using

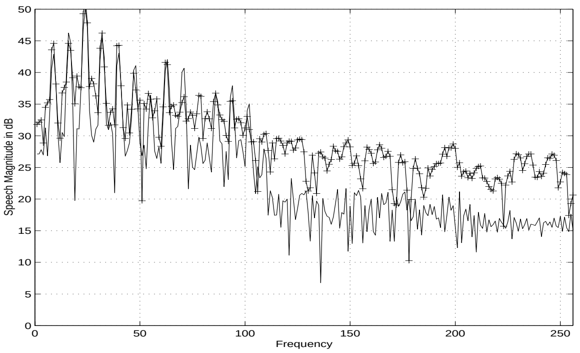

Figure 3: The spectrum of a speech sentence in dB, from the NTIMIT database before warping versus the spectrum of the same sentence , represented by + symbols, after warping.

512-point FFT of that frame is computed. Then a bank of bandpass filters is centered at each peak of the short-term spectrum of the speech signal with an appropriate bandwidth. It was found by experiments that the first 20 peaks are enough to obtain a consistent $\beta(n)$. Once we obtained the AM-FM signal, the instantaneous frequency calculation is accomplished via the **SEOSA** described above. Then the average fundamental frequency is calculated using Equation (12). $\beta(n)$, consequently, can be found by solving Equation (13). Then the iterative algorithm described in section (3) is used for re-estimating $\beta(n)$ to make the quasi-harmonics more periodic. The more iterations that we use, the more our algorithm is more consistent; consequently $\beta(n)$ is more accurate and the warped signal is more periodic.

Fig. (3) presents two spectra; the spectrum of an original speech sentence, i.e., before warping versus the spectrum of the same sentence, represented by + symbols, after warping. It is clear from Fig. (3) that the high-frequency harmonic peaks of the warped spectra would be approximately 10 dB higher than the original spectra. That is, the inclusion of the warping with the **SEOSA** enhances the voiced part of the speech sound and consequently enhances the speech signal. This enhancement makes the **SEOSA** more robust when used in different noisy-speech processing areas, i.e., makes the algorithm less sensitive to noise and more accurate in extracting different speech parameters from the speech signal.

5 Conclusion

In this paper, we have presented a novel technique of speech enhancement via the smoothed energy operator separation algorithm **(SEOSA)**. The enhancement algorithm was presented in the framework of the AM-FM speech modulation model. Experiments have shown that the harmonic peaks of the warped spectra would be approximately several dB higher than the original spectra. This gain is useful in extracting the speech model parameters more correctly in the presence of noise.

One possibility for the future research is to improve the enhancement algorithm described above by applying comb filtering to the voiced part of the speech signal. This should be helpful to make our system more robust to noise. Another possibility is to test this warping technique on all speech applications that use the **(SEOSA)** described in [2]-[7].

We are currently continuing the effort towards the use of our novel technique when combined with comb filtering as the front-end of an automatic speech recognition system to improve its performance when used to recognize noisy speech signals.

References

[1] James Kaiser, “On a Simple Algorithm to Calculate The Energy of a Signal”, Proc. ICASSP, pp. 381-384, 1990.

[2] P. Maragos, T. Quatieri and J. Kaiser, “Speech Nonlinearities, Modulations and Energy Operators”, Proc. ICASSP, pp.421-424, 1991.

[3] P. Maragos, T. Quatieri and J. Kaiser, “On Amplitude and Frequency Demodulation Using Energy Operators”, IEEE Trans. on Signal Processing, Vol. 41, No. 4, pp. 1532-1550, April 1993.

[4] P. Maragos, J. Kaiser and T. Quatieri, “Energy Separation in Modulations with Application to Speech Analysis”, IEEE Trans. on Signal Processing, Vol. 41, No. 10, October 1993.

[5] A. Bovik, P. Maragos, T. Quatieri, “AM-FM Energy Detection and Separation in Noise Using Multi-band Energy Operators”, IEEE Trans. on Signal Processing, Vol. 41, No.12, December 1993.

[6] A. Potamianos and P. Maragos, “A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation”, Signal Processing, Vol. 37, pp. 95-120, May 1994.

[7] A. Potamianos, “Speech Processing Applications Using an AM-FM Modulation Model”, Ph.D. dissertation, Harvard University, 1995.

[8] M. Ramalho and R. Mammone, “A New Speech Enhancement Technique with Application to Speaker Identification”, Proc. ICASSP, pp. I.29-32, 1993.

[9] M. Ramalho and R. Mammone, “The Pitch Mode Modulation Model and its Application in Speech Processing”, The Kluwer Academic Series in Engineering and Computer Science: Modern Methods of Speech Processing, pp. 377-400, 1995.