COMBATING NONLINEAR TELEPHONE CHANNEL-NOISE USING THE MULTIBAND AM-FM MODEL

Hesham Tolba *Douglas O'Shaughnessy*

INRS-Télécommunications, Université du Québec
16 Place du Commerce, Verdun (Île-des-Soeurs),
Québec, H3E 1H6, Canada
{tolba, dougo}@inrs-telecom.uquebec.ca

ABSTRACT

This study presents a novel technique to enhance telephone speech signals. This technique is based on the Amplitude and Frequency Modulation (AM-FM) model, which represents the speech signal as the sum of N successive AM-FM signals. Based on a least-mean-square error criterion, each AM-FM signal is modified using an iterative algorithm in order to compensate for the deformation of the signal caused by the nonlinear telephone channel. These modified signals are then combined in order to reconstruct the enhanced speech signal. Experiments were conducted using speech signals extracted from the NTIMIT database. Such experiments demonstrate the ability of the algorithm for speech enhancement, in terms of a comparison between the original and synthesized speech and informal listening tests.

1. INTRODUCTION

The problem of processing band-limited (telephone) speech signals has received considerable attention in order to improve the subjective speech quality, especially with the significant increase in speech applications over telephone lines. Several techniques have been proposed in the literature to solve this problem such as [1]; however such techniques suffer from several drawbacks and are not able to recover from all the distortions introduced by the telephone channel [2]. In this study, we present a novel speech enhancement technique, which is based on a multiband analysis of the speech signal using the AM-FM model. We show through experiments that using such a model, we are capable of enhancing the speech signal affected by the telephone channel by modifying the AM-FM signals obtained using the multiband AM-FM analysis.

This novel technique enhances the telephone speech signals using an iterative approach. In this proposed approach, the AM-FM model [7] and a multi-band analysis scheme [9] are applied to the speech signal to extract the spectrum of each AM-FM signal at the successive resonances of the speech signal. This process is applied to the TIMIT training speech data in order to extract the isolated AM-FM signals, which are used to compute a weighting function ω_i, that is used to modify the corresponding NTIMIT AM-FM signals based on a mean-square error (MSE) criterion. The obtained weighting functions serve in the combating the deformation caused by telephone channels. The estimated signals are then used to regenerate the enhanced speech. To reconstruct the speech signal, the modified AM-FM signals are accumulated together in the time domain on a frame-by-frame basis. Then these speech frames are concatenated in order to synthesize the enhanced speech signal.

The outline of this paper is as follows. In sections 2 and 3, an overview of the effects of the nonlinear-telephone channels on speech transmission and the AM-FM Model, respectively, is given. Then in section 4, we describe the proposed approach for speech enhancement and the system designed for such an approach. Experimental results that demonstrate the effectiveness of our algorithm are presented in section 5. Finally, in section 6 we conclude and discuss our present and future work.

2. BACKGROUND

Many signal degradations occur in the transmission of voice from one user to another over a channel. Main sources of such degradations include: distortion (such as harmonic, inter-modulation and A/D/A distortion), echo, bandwidth limitation and network noise (such as thermal noise, shot noise, low-frequency noise, single-frequency noise and crosstalk). In this study, we focus on speech affected by channel distortion sources, which include most of the above sources of degradations, that is, telephone and microphone distortions. Studying telephone speech shows that although it has high SNR, it is of a low quality and difficult to recognize due to its bandwidth limitation, handset and connection quality variability and sometimes background noise. The main sources of telephone line distortions are: burst (impulse noise), hum, echo, frequency translation, modulation distortion, unknown channel gain, added low-frequency tones, breath intake and release, A/D/A distortion which includes quantizing and aliasing distortion and additive stationary noise. Some of these noises are additive in the spectral domain, while the others are convolutional.

On the other hand, distortion induced by microphones can be summarized into three main sources: convolutional noise, additive noise and reverberation. Convolutional noise is generated due to the fact that microphones act as linear filters on the speech signal and account for different degrees of spectral slope depending on the microphone characteristics. Microphone sensitivity to the background noise corresponds to additive perturbations of the recording speech. Furthermore, as the microphone-to-talker distance is often different, gain variation and reverberation is also observed. That is why automatic speech recognition systems perform

poorly when different microphones are used for training and testing. In [5], it was found that the error rate increased by 62% when alternate microphones than that used in the training phase were used for testing.

Quatieri et al. in [3] and [4] studied the effects of the nonlinear distortions from the carbon-button microphone and telephone channels. Comparison of TIMIT and NTIMIT speech segments in the above-mentioned studies shows that the nonlinear channel effects introduce *phantom formants* at sums, differences and multiples of the original formant frequencies. The amount of such phantom formants was found to be on the order of 10% and 15% between F1 and F2 and 16% and 18% between F2 and F3 for male and female speakers, respectively [4]. Besides, resonance bandwidths can be broadened or narrowed depending on the order of nonlinearity. In this paper, we propose an approach to solve such problems. As mentioned above, this approach was built in the AM-FM model framework described in section 3.

3. THE AM-FM MODEL FRAMEWORK

3.1. Analysis-Synthesis

The *AM-FM Model* introduced in [7] was found to be efficient, simple and accurate in representing speech signals. In such a model, the speech signal, $s(t)$, is modeled as a sum of AM-FM signals as follows:

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

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. The instantaneous frequency $f_{inst}(t)$ of each component of the speech signal is defined as:

$$f_{inst_i}(t) = \frac{d}{dt}[\phi_i(t)]. \tag{2}$$

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

$$s(n) = \sum_{i=1}^{N} a_i(n) cos[2\pi f_i n T + \theta_i] \tag{3}$$

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

$$\phi_i(n) = \overline{\omega}_i n + \theta_i \qquad \forall i, \tag{4}$$

where $\overline{\omega}_i$ is the average peak frequency for the window time interval and θ_i is the phase offset.

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 must be performed. Then these parameters can be estimated for each resonance using an energy tracking operator such as *SEOSA*.

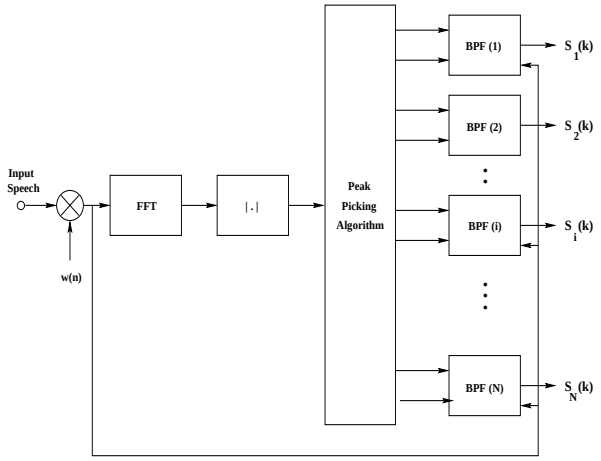

Figure 1: Block Diagram of the multiband AM-FM-based SEOSA analysis speech enhancement system.

3.2. Multi-Band AM-FM Demodulation

By the selection of an appropriate bandpass filter, isolation of an AM-FM signal component could be possible [9]. This is due to the fact that the wideband FM signal could be restricted to a limited-band signal without losing 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. 400 Hz was found in [8] to be a reasonable value for the bandwidth of such a filter in order to avoid the effects of neighboring formants.

Resonance isolation is performed using a bank of bandpass filters. These filters are implemented using a truncated, discretized bandpass FIR filter (BPF) with impulse response

$$g_\omega(n) = g(n)\, \omega_H(n), \tag{5}$$

where $\omega_H(n)$ is the Hamming window function

$$\omega_H(n) = \begin{cases} 0.54 - 0.46cos\left(2\pi \frac{n}{L-1}\right), & 0 \le n \le L-1, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

which is used to modify and truncate the ideal impulse response $g(n)$ given by:

$$g(n) = \frac{sin[\omega_H(n - L/2)]}{\pi(n - L/2)} - \frac{sin[\omega_L(n - L/2)}{\pi(n - L/2)]}, \tag{7}$$

where ω_H and ω_L define the low and high edges of the passband and $L/2$ is the delay required for causality.

4. SPEECH ENHANCEMENT SYSTEM

The speech enhancement system proposed in this paper is based upon a multi-band analysis described in section 3.2 applied to the speech signal modeled using the AM-FM speech model described in section 3. The idea of this proposed approach is to modify the spectrum of the N successive high-frequency bands of the bandpass filtered narrow-band speech signal in order to restore the missing frequency bands. Such a modification is obtained through an iterative adaptation technique based on a least square estimate during a training phase. The coefficients obtained in such a training are then used in order to modify the narrow-band speech spectrum. These spectrums, in turn, are then added together in order to synthesize the wide-band speech signal. Such a reconstruction is performed in the time domain on a frame-by-frame basis.

4.1. Spectral Estimation

The AM-FM speech analysis window used in this study consists of three parts: the first and third parts are half Hamming windows, whereas the second part is a rectangular window. This window function is given by:

$$\omega_T(n) = \begin{cases} \omega_H(n), & 1 \leq n \leq L_\omega, \\ 1, & L_\omega + 1 \leq n \leq 2L_\omega, \\ \omega_H(n - 2L_\omega + L_\omega), & 2L_\omega + 1 \leq n \leq N_\omega, \end{cases} \quad (8)$$

where $\omega_H(n)$ is the Hamming window function (Eq. 6, length = $2L_\omega$), N_ω is the window length and $L_\omega = N_\omega/3$. The frame size, N_ω, is chosen equal to 30 ms, with 10 ms displacement. The AM-FM analysis window is applied to an equal number of samples (160 samples) from the past speech frame, the present speech frame and the future speech frames, respectively. Then, each windowed speech frame is transformed into the frequency domain using a 1024-point FFT. Then the spectrum of the N AM-FM signals is computed for each peak in the Fourier spectrum by isolating each peak using a filter bank of the bandpass filters (described in section 3) centered around each peak. The bandwidth for such bandpass filters was chosen to be approximately 400 Hz as mentioned in section 3.

Once we isolate each AM-FM signal, the spectra for these successive resonances are then modified using an adaptive filter. The coefficients of such a filter are obtained during a training phase, using a least-mean-square (LMS) adaptive algorithm given the IFFT of the spectral obtained from the narrow-band speech signal, $s_{k,i}(n)$, as the input signal to such a filter and the IFFT of the spectral obtained from the broad-band speech signal, $s_{k,i}(n)$, as the desired signal. The output of such a filter is given by:

$$\hat{s}_{k,i}(n) = \sum_{j=1}^{M} \omega_i(j) s_{k,i}(n-j), \quad (9)$$

where $\hat{s}_{k,i}(n)$ is the estimate of the recovered AM-FM signal around the i^{th} peak for the k^{th} frame, M represents the filter's length and ω_i represents the filter coefficients for such a peak. In our experiments, we used filters consisting of 41 taps.

4.2. Speech Synthesis

Once we have obtained the modified AM-FM signals based on the above mentioned LMS algorithm, the synthesis of the broad-band speech is performed by the summation of the modified AM-FM signals, $\hat{s}_{k,i}(n)$, in order to obtain the speech signal for each frame k, $\hat{s}_k(n)$. Then, for each of these signals, only the second 160 samples, which represent the current frame, are selected, then concatenated together in order to generate the broad-band speech signal.

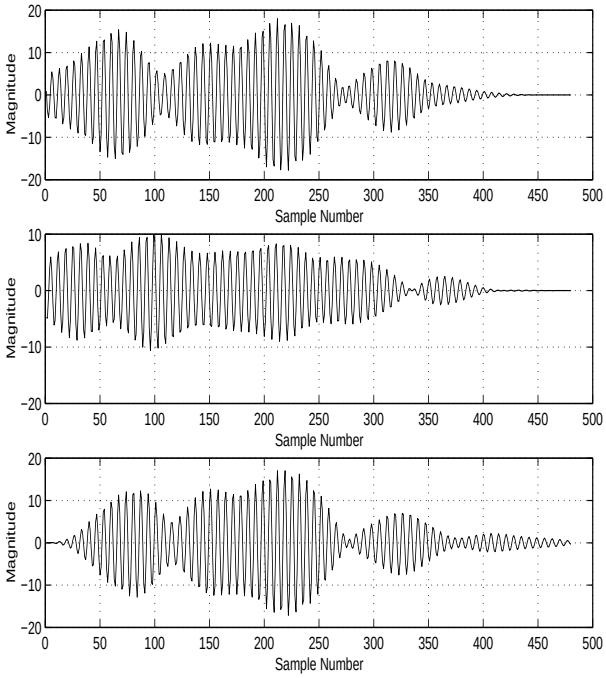

Figure 2: Comparison of a single AM-FM time waveform of (f_c=2297 Hz; BW ≈ 400 Hz) (a) the original broad-band speech, (b) the narrow-band speech and the (c) the estimated speech.

5. EXPERIMENTAL RESULTS

The speech corpus for this experiment is a subset of the NTIMIT database [10]. The NTIMIT database is the telephone-bandwidth version of the widely-used TIMIT database, which was sampled at 16 kHz. Besides band limitation, several kinds of distortion can be found in the NTIMIT database, such as broadband noise, band-limiting, low frequency hum, crosstalk, dial pulses, shot noise and sharp pulses. Utterances from both the TIMIT and NTIMIT databases were selected in order to train our algorithm to obtain the weighting coefficients ω_i. However, sentences outside the training set were used for the evaluation of the algorithm. Such experiments demonstrate the ability of the algorithm for speech enhancement, in terms of a comparison between the original and synthesized speech and informal listening tests.

Fig. 2 shows a typical example of an AM-FM signal at f_c=2297 Hz that, using the previously-mentioned MSE criterion, we restore the broadband AM-FM signal from the narrow-band one around that peak. The more training data that we use, the more our algorithm is accurate and our estimation is better.

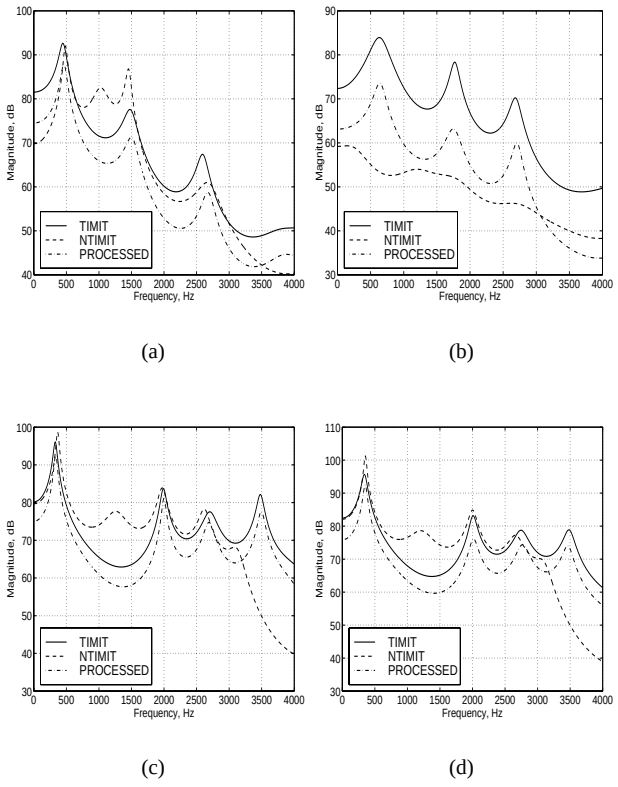

Figure 3: Comparison between original, bandlimited and enhanced LPC spectra.

Figs. 3.a-3.d illustrate typical examples of LPC spectra of speech frames of 30-msec length selected from the TIMIT database, the LPC spectra of the corresponding frames from the NTIMIT database and the LPC spectra of the enhanced speech frames using the proposed approach. These figures indicate that we have eliminated most of the distortion introduced by the telephone channel. The amount of accuracy of such an approach depends on several factors such as: the order of the nonlinearity introduced by the channel, the amount of training data used to design the adaptive filters (in other words, how accurate is the adaptive filter), the number of the AM-FM signals, N, consequently the number of filters that have been used in the analysis and synthesis of the speech signal and the associated bandwidths (i.e., the design of the filter bank that is used for analysis and synthesis of the speech signal).

6. CONCLUSION

In this paper, we have presented a novel speech enhancement approach. The enhancement algorithm was presented in the framework of the AM-FM speech modulation model. We have examined several possibilities for the enhancement of the distorted telephone-speech signal in terms of the parameters obtained from the AM-FM speech model. The extracted parameters for each AM-FM signal, extracted based on a LMS criteria using both the original speech (TIMIT) and the telephone speech (NTIMIT) signals, are used to find the characteristics of the telephone channel within the different bands of the AM-FM signals. These characteristics are then used to remove the nonlinear distortion occurred to the speech signals by the telephone channels. In these respects, an algorithm has been built and tested on telephone speech. Comparisons between the original and synthesized speech indicated that such an algorithm is able to eliminate the phantom formants and the deformation of the resonance bandwidths, produced by the telephone channels, for different speakers. Moreover, informal listening tests indicate that the subjective quality of bandlimited speech is enhanced using our proposed algorithm.

One possibility for future research is to improve the enhancement algorithm described above by using a set of contiguous bandpass filters, whose passbands increase in width with increasing frequency. That is, applying analysis which is similar to the frequency analysis made by the ear that is characterized by the so-called *critical bands*.

7. REFERENCES

[1] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech (RASTA-PLP)", Proc. EUROSPEECH–91, pp. 1367–1370, 1991.

[2] P. Moreno and R. Stern, "Sources of Degradation of Speech Recognition in the Telephone Network", Proc. ICASSP–94, pp. I.109–I.112, 1994.

[3] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary and B. A. Carlson, "The Effects of Telephone Transmission Degradations on Speaker Recognition Performance", Proc. ICASSP–95, pp. 329–332, 1995.

[4] T. F. Quatieri, D. A. Reynolds and G. C. O'Leary, "Magnitude-Only Estimation of Handset Nonlinearity with Application to Speaker Recognition", Proc. ICASSP-98, 1998.

[5] H. Murveit and J. Butzberger and M. Weintraub, "Performance of SRI's DECIPHER Speech Recognition System on DARPA's CSR Talk", Proc. DARPA Workshop Speech and Natural Language, pp. 410–414, 1992.

[6] James F. Kaiser, "On a Simple Algorithm to Calculate the Energy of a Signal", Proc. ICASSP-90, pp. 381-384, 1990.

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

[8] P. Maragos, J. F. Kaiser and T. F. Quatieri, "Energy Separation in Modulations with Application to Speech Analysis", IEEE Trans. on Signal Processing, Vol. 41, No. 10, pp. 3024-3051, October 1993.

[9] A. Bovik and P. Maragos and T. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multi-band Energy Operators", IEEE Trans. on Signal Processing, SP–41(12), pp. 3245-3265, December, 1993.

[10] Charles Jankowski, Ashok Kalyanswamy, Sara Basson and Judith Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", Proc. ICASSP-90, pp. 109-112, 1990.