AM-DEMODULATION OF SPEECH SPECTRA AND ITS APPLICATION TO NOISE ROBUST SPEECH RECOGNITION

Qifeng Zhu and Abeer Alwan

Department of Electrical Engineering, UCLA
Los Angeles, CA 90095

ABSTRACT

In this paper, a novel algorithm that resembles amplitude demodulation in the frequency domain is introduced, and its application to automatic speech recognition (ASR) is studied. Speech production can be regarded as a result of amplitude modulation (AM) with the source (excitation) spectrum being the carrier and the vocal tract transfer function (VTTF) being the modulating signal. From this point of view, the VTTF can be recovered by amplitude demodulation. Amplitude demodulation of the speech spectrum is achieved by a novel nonlinear technique, which effectively performs envelope detection by using amplitudes of the harmonics and discarding inter-harmonic valleys. The technique is noise robust since frequency bands of low energy are discarded. The same principle is used to reshape the detected envelope. The algorithm is then used to construct an ASR feature extraction module. It is shown that this technique achieves superior performance to MFCCs in the presence of additive noise. Recognition accuracy is further improved if peak isolation [1] is also performed.

1. INTRODUCTION

The acoustic speech signal contains information about both the excitation (source) signal and the vocal tract transfer function (VTTF). There are applications where it is important to accurately estimate the VTTF and to discard variations due to changes in fundamental frequency or pitch. For example, the VTTF is often used in feature extraction for Automatic Speech Recognition (ASR). Linear Prediction Coding (LPC) analysis estimates the VTTF with an all-pole model. However, LPC based features (for example LPCC) are vulnerable to background noise. Similarly, an FFT spectrum or a smoothed version of it will be sensitive to background noise. In this paper, we introduce a noise robust technique for estimating the envelope of the speech spectrum, which contains information on the VTTF. The technique resembles amplitude demodulation in the frequency domain.

The use of the term “modulation” in this paper is different than that used by others. For example, “modulation spectrum” [2] [3] uses low-pass filters on the time trajectory of the spectrum to remove fast-changing components. In [4], the authors model speech waveforms as amplitude and frequency modulated (AM-FM) signals where formant frequencies are the frequencies of the carriers.

In Section 2, we introduce the theory behind the proposed algorithm, and compare it to linear envelope detection. In Section 3, we evaluate the technique as a front end for an HMM based automatic speech recognition system at different SNRs.

2. HARMONIC DEMODULATION

2.1. Theory of Speech Production and Amplitude Modulation (AM)

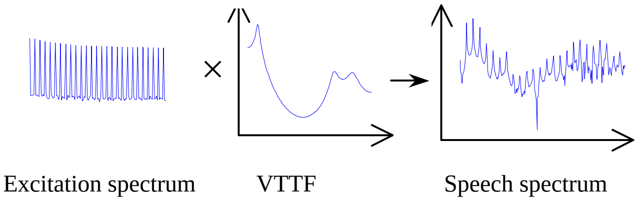


Figure 1: The linear source-filter model of speech production in the frequency domain. The speech spectrum is the result of the multiplication of the excitation spectrum with the spectrum of the vocal tract transfer function (VTTF). The excitation spectrum in this example in harmonically related (voiced speech). The x-axis is frequency.

The linear source-filter model [5] of speech production views the speech waveform as the result of convolution between the excitation signal (which is either quasi-periodic, noise like, or a combination of the two) and the impulse response of the vocal tract transfer function (VTTF). In the frequency domain, the speech spectrum is the result of multiplication of the source (excitation) spectrum and the VTTF, as shown in Figure 1. For voiced signals the excitation spectrum is harmonic.

The speech spectrum can also be viewed as a result of amplitude modulation (AM) in the frequency domain with the source (excitation) spectrum being a carrier and the vocal tract transfer function (VTTF) being the modulating signal. Typically, amplitude modulation refers to modulation in the time domain, as shown in Figure 2, where the carrier in this example is a high frequency sinusoid and the modulating signal is a slowly-varying signal. When the carrier is noise, the noise spectral envelope is modulated in a similar way.

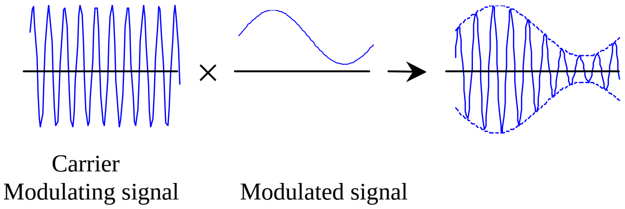


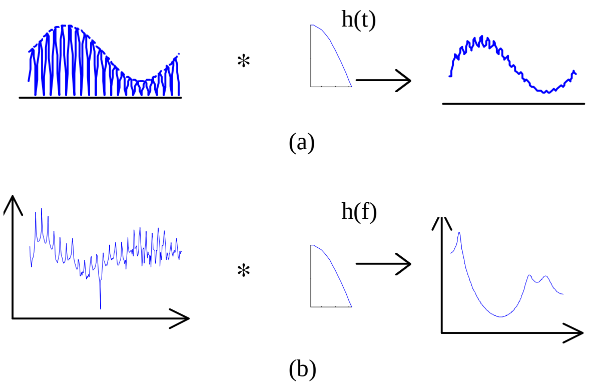
Figure 2: Amplitude modulation. The modulated signal is the multiplication of the carrier and the modulating signal in the time domain.

The modulation framework is a simplified speech production model. For example, in speech production, the carrier is not strictly sinusoidal, and the amplitude of the carrier is not constant.

2.2. Demodulating the Speech Signal

Our goal is to estimate the vocal tract transfer function and remove any pitch-related information. This is related to demodulating the speech spectrum in the frequency domain.

Coherent demodulation [6], used in AM radio for example, requires recovering the carrier signal. Incoherent demodulation, on the other hand, involves envelope detection using a rectifier and low-pass filter. Figure 3(a) illustrates the incoherent demodulation process in the time domain after the modulated signal has been full-wave rectified. We will adopt a similar strategy but perform it in the frequency domain, as shown in Figure 3(b), where the resultant spectrum is “harmonically demodulated”.



(a)

(b)

Figure 3: (a) Envelope detector for AM demodulation. First, full wave rectification is applied, and then the signal is convolved with the impulse response of a low-pass filter. The x-axis here is time. (b) Similar process to (a), except that it is done in the frequency domain by using the magnitude of a speech spectrum. The x-axis here is frequency.

2.2.1. Linear and Non-linear Envelope Detection

The process in Figure 3(b) can be implemented by linear convolution between the speech spectrum and the characteristic of the low-pass filter. Convolution is performed as the superposition of the convolutions between every point in the DFT spectrum and the response of the low-pass filter. The equation which is used to compute the convolution is:

$$S(k) * h(k) = [\sum_i S(i) d(k-i)] * h(k) = \sum_i [S(i) d(k-i) * h(k)]$$

S(k) is the discrete speech spectrum, and *h(k)* is the characteristic of the low-pass filter in the frequency domain.

This process is illustrated in Figure 4. Figure 4(a) is a simplified representation of the input speech spectrum. The highest spikes are the harmonic peaks. The other samples represent inter- harmonic frequency components which maybe due to noise or other factors. The inter-harmonic peaks are simplified in this example to either 0 or half the amplitude of the harmonic peaks. The envelope of the harmonic peaks is flat in this example. We will illustrate how to recover this envelope. Figure 4(b) shows an example of the characteristic of the low-pass filter. We first compute the convolution between each point in the input spectrum with the characteristic of the low-pass filter, as explained earlier. Figure 4(c) shows the results of the convolution process of every point in Figure 4(a) and Figure 4(b).

Finally, the convolution of the whole spectrum with the filter response is obtained as the superposition of the results in Figure 4(c). The result is shown in Figure 4(d) as the top solid line.

One problem of this linear envelope detection is that it is vulnerable to inter-harmonic components, hence it would not be robust to background noise.

Alternatively, one can perform envelope detection in a way that is less susceptible to inter-harmonic components. We achieve this by a non-linear technique, hereafter refer to as NELD, which effectively estimates the envelope by focusing only on the harmonic peaks. Instead of computing the superposition of the results in Figure 4(c), we compute the maximum with the equation:

$$\underset{i}{Max} [S(i) d(k-i) * h(k)]$$

The result of demodulation using this non-linear technique for the same input spectrum is shown in Figure 4(d), as the line with triangles. We can see that the result is smoother than the envelope detected linearly.

2.2.2. Robustness Analysis

As mentioned in the last section, linear envelope detection is not noise robust. We will illustrate this with an example.

Figure 5(a) shows the same speech spectrum as Figure 4(a), except that one inter-harmonic point has increased in amplitude due to background noise (indicated by an arrow in Figure 5(a)). With the linear envelope detection technique, this change will affect the output, as seen in the top line in Figure 5(d). If the non-linear technique is used, the output is not affected, as seen in Figure 5(d) in the line with triangles.

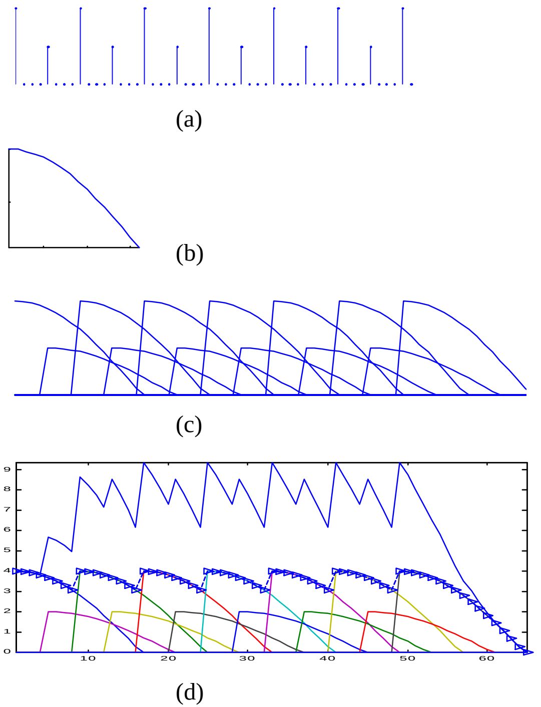


Figure 4: Envelope detection. (a) A simplified speech spectrum. (b) The response of the low-pass filter for envelope detection. (c) Results of the convolution between every point in (a) and (b). (d) The envelope estimated by linear demodulation (superposition) is shown as the solid line in the upper part of the figure. Also shown is the envelope detected using NLED (line with triangles).

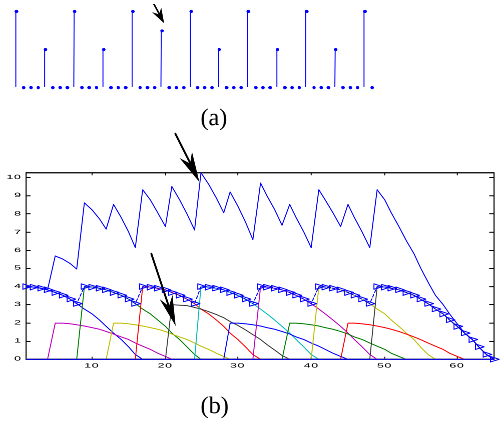


Figure 5: (a) A simplified speech spectrum (the modulated signal) with additive noise at one point indicated by an arrow. (b) The demodulated signal. The top solid line is the envelope detected by linear convolution while the lower line (with triangles) results from NLED.

Obviously, if the increase results in amplitudes that are larger than the harmonic peaks then both the linear and non-linear techniques will be affected.

The noise robust capability of the NLED technique can be understood by observing that frequency bands with low energy, such as inter-harmonic frequencies are more susceptible to noise.

3. USING HARMONIC DEMODULATION IN ASR

To determine whether harmonic demodulation can be used as a noise robust feature extraction method in ASR, we used it in computing Mel Frequency Cepstral Coefficients (MFCC) and performed recognition experiments. MFCCs are the result of performing a DCT on log spectral estimation obtained with a critical bandwidth like non-uniform filter bank. In our evaluations, MFCCs are calculated using the log spectral estimate of the speech signals after harmonic demodulation. A block diagram illustrating how harmonic demodulation is used is shown in Figure 6.

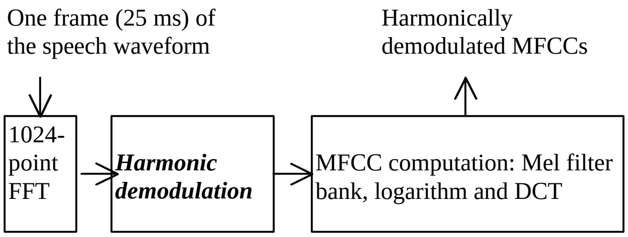


Figure 6: Block diagram of the harmonically demodulated MFCCs.

3.1 Implementing the HDMFCC

The harmonically demodulated MFCCs (HDMFCC) are either computed using linear demodulation or using the non-linear demodulation technique introduced in this paper. Speech is sampled at 12.5 kHz and 25 ms frames, overlapped by 15 ms, are obtained with a Hamming window. Pre-emphasis is used. For each frame, a 1024 point FFT is computed, and only half the points are used because of the FFT symmetry. This corresponds to the frequency range between 0 to 6250 Hz. The characteristic of the low-pass filter used in envelope detection is shown in Figure 7. The width of the filter is 43 points which corresponds to 525 Hz, and the magnitude is above 0.8 for about 210 Hz. The characteristic of the filter was optimized to achieve high accuracy in speech recognition experiments.

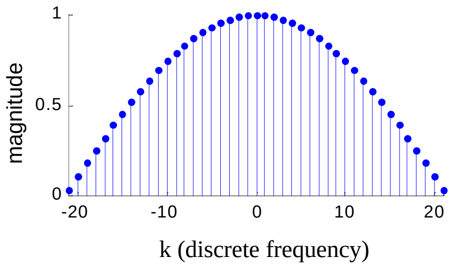


Figure 7: Characteristic used in envelope detection.

The envelope obtained from NLED can be further reshaped for better noise robustness using the same principle of avoiding frequency regions with low energy. Envelope values are only considered if they are above a certain threshold. The threshold is set empirically to be half the mean amplitude of the 512 point FFT speech spectrum before envelope detection.

3.2 Recognition Experiments

In these experiments, a Hidden Markov Model (HMM) based system (HTK2.1) was used for an isolated digit recognition task using the TI46 database. For each digit, one HMM with 4 states and 2 mixtures is trained from 160 utterances spoken by 16 talkers (8 males and 8 females). Training includes 2 steps of Maximum Likelihood (ML) and Expectation Maximization (EM) with 4 iterations each. A Viterbi algorithm is used for recognition using 960 different utterances. Training is done with clean signals, and recognition with noisy signals (speech with additive speech shaped noise) at different SNRs.

The following features were used in the experiments. 1) MFCCs, 2) HDMFCCs with linear demodulation 3) HDMFCCs with non-linear demodulation 4) MFCCs with peak-isolation [1] (referred to as to MFCCP), 5) MFCCs with non-linear demodulation and envelope reshaping, and 6) HDMFCCs with non-linear demodulation, envelope reshaping, and peak-isolation. Results are shown in Figure 8.

As seen in the figure, as the SNR decreases, demodulation, whether linear or non-linear, improves recognition performance. In addition, NLED, envelope reshaping together with a process that enhances peaks in the spectrum improve dramatically recognition performance without a significant increase in computational cost. For example at SNR of 3 dB, the recognition accuracy is 38 percent for MFCCs, versus 78 percent for the proposed algorithm with peak-isolation.

4. SUMMARY AND CONCLUSION

In this paper, a novel algorithm that resembles amplitude demodulation in the frequency domain is introduced using a non-linear envelope detection (NLED) technique. The NLED relies on the amplitudes of the harmonics and avoids inter-harmonic valleys. The algorithm differs from linearly smoothing the speech spectrum or deconvolution of the source and vocal tract impulse response. This technique is noise robust since envelope detection does not take into account frequency regions of low signal energy. The same principle is used to reshape the envelope after it is detected. The algorithm is then used to construct an ASR feature extraction module. It is shown that this technique achieves superior performance to MFCCs in the presence of background noise. Recognition accuracy is further improved if peak isolation [1] is also performed.

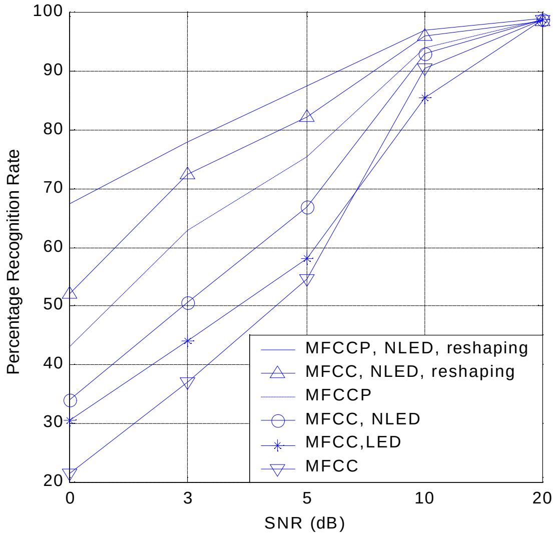


Figure 8: Recognition results with additive speech shaped noise at different SNRs.

Acknowledgments

Work supported in part by NSF and funding from HRL through the UC-MICRO program.

5. REFERENCES

1. Strope, B. and Alwan, A. 1997. “A model of dynamic auditory perception and its application to robust word recognition”, IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 5, p. 451-464.
2. Kanedera, N. Hermansky, H. Arai, T. “On properties of modulation spectrum for robust automatic speech recognition”, Proc. ICASSP '98, vol.2. p.613-16.
3. Greenberg, S. and Kingsbury, B.E.D. “The modulation spectrogram: in pursuit of an invariant representation of speech”, Proc. ICASSP '97, vol.3, p.1647-50.
4. Potamianos, A. and Maragos, P. “Speech analysis and synthesis using an AM-FM modulation model”, Speech Communication, vol.28, (no.3), 1999, p.195-209.
5. Fant Gunnar. “The Acoustic theory of speech production”. S’Gravenhage, Mouton, 1960.
6. Haykin, Simon S., “Communication systems”, New York, Wiley, c1978.