BANDWIDTH EXTENSION OF TELEPHONE SPEECH USING FRAME-BASED EXCITATION AND ROBUST FEATURES

Ismail Uysal, Harsha Sathyendra, John G. Harris

Computational NeuroEngineering Laboratory, The University of Florida
Gainesville, Florida, 32611, USA
phone: + (1) 352-392-2626, fax: + (1) 352-392-0044, email: uysal@ufl.edu, hsathyen@ufl.edu, harris@cnel.ufl.edu
web: www.cnel.ufl.edu

ABSTRACT

The standards that are still in use for telephone communications since the 1950s limit the information bandwidth to 300-3400Hz. However, in normal conversational speech, the frequency content is mainly between 0-8000Hz. This constraint degrades not only the sound quality but also the intelligibility of the transmitted signal. Instead of modifying the present telecommunication infrastructures, which would cost billions of dollars, many researchers have been studying more efficient methods to increase the quality of telephone speech. This paper develops an innovative solution to bandwidth extension, which is based upon the Linear Source Filter Model that breaks speech up into two parts: the excitation and the spectral envelope. Novel approaches are used to extend the frequency information for both parts. This algorithm particularly emphasizes low frequency reconstruction without neglecting high frequencies. Furthermore, different feature sets to model the spectral envelope are employed for better performance under noisy conditions.

1. INTRODUCTION

One of the most common ways to analyse speech is to use a Linear Source Filter Model (LSFM) that results in passing a glottal excitation signal through a linear time varying all-pole filter [1]. This linear filter models the vocal tract resonations, whereby the filter's magnitude response is defined as the spectral envelope. Glottal excitation resembles either an impulse train or spectrally flat white noise, depending on whether the speech signal is voiced or unvoiced, respectively. Speech can be then modelled as the convolution of the glottal excitation and the vocal tract filter. The speech signal is inherently non-stationary and thus usually broken up into small time frames in order to approximate stationary behaviour. These frames are further broken up into excitation and spectral envelope parts.

Narrowband (NB) speech is band-limited to 300-3400Hz. In order to increase the quality and intelligibility of speech, bandwidth extension (BWE) is used to extend NB speech to wideband (WB) speech, 0-8000 Hz. However, after extensive experimental trials, it is found that there are only marginal differences in sound quality and intelligibility between halfband (HB) speech 0-4000Hz, and WB speech, whereas there are substantial differences between HB and NB speech (telephone speech). Therefore, the work presented in this paper concentrates on extending NB to HB speech. This notion deviates from the usual aspects of BWE, which maps HB to WB speech. To keep with convention, from this point on, HB speech will be referred to as WB speech. As seen in Fig.1, after extensions of both the spectral envelope and the excitation, the resulting speech will be called reconstructed wideband (RWB) speech.

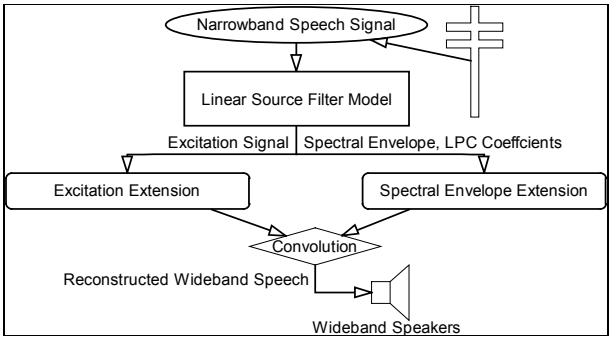

Fig. 1. Block Diagram of Bandwidth Extension Algorithm

2. EXTENSION OF EXCITATION SIGNAL

According to LSFM, speech can be broken up into two parts: the excitation and the spectral envelope. In order to attain high quality WB speech, both parts have to be extended. This section concentrates on the extension of the excitation signal. In order to isolate the problem, this part of the algorithm utilizes the WB spectral envelope and extrapolates the NB excitation signal.

2.1 Previous Excitation Notions

The formal versions of excitation extension algorithms state that the excitation signal $U_g(k)$ is nearly spectrally flat [2,5,6]. When considering a strictly band-limited speech input signal, the spectrally flat excitation assumption only holds for unvoiced frames. For voiced frames, the excitation signal consists of impulsive components placed at pitch harmonics.

The main reason for considering the excitation to be spectrally flat for all frames is for simplicity, whereby the extension is done via a simple modulation as seen below:

$$U_{eb}\left(e^{j\Omega}\right) = U_{nb}\left(e^{j(\Omega-\Omega_M)}\right) + U_{nb}\left(e^{j(\Omega+\Omega_M)}\right)$$

where,

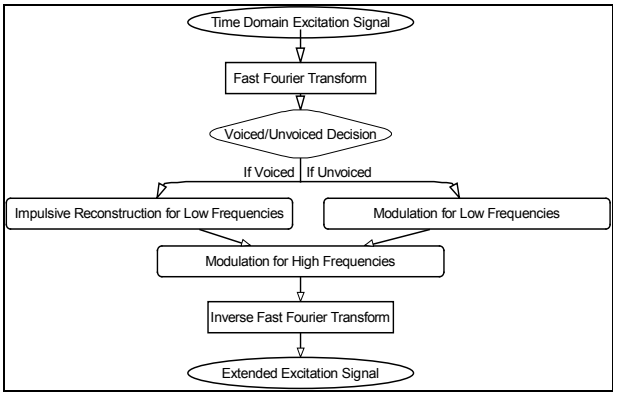

Fig. 2. Block Diagram of Excitation Extension Algorithm

$$\Omega_M \approx \Omega_{3.4} = 2\pi \cdot \left(\frac{3.4KHz}{f_s} \right)$$

where f_s is the sampling frequency.

For voiced frames the harmonic structure of the excitation signals contain impulses at multiples of the fundamental (pitch) frequency of the speech segment. Therefore, reconstructed excitation must convey comprehensive pitch information even though the first few harmonics of pitch frequency may be lacking in the NB speech.

2.2 Enhancing Previous Excitation Algorithm

For this frame-based excitation extension (F-BEE), the speech signal is first broken up into frames and classified as voiced and unvoiced frames. Frames are labelled as such, via the spectral flatness measure (SFM) [3].

$$SFM = 10\log_{10}\left(\sqrt[1/N]{\prod_{k=1}^{N}|X(k)|} \right) / \left(\frac{1}{N}\sum_{k=1}^{N}|X(k)| \right)$$

where X(k) is the FFT of a single frame which has N samples. For speech, the SFM results in values around 0dB for voiced frames and around -60dB for unvoiced frames. As seen in Fig.2, for unvoiced frames, a simple modulation is adequate. However, for voiced frames, an impulsive reconstruction is required for low frequencies where more an impulsive behaviour than spectral flatness is observed. The first step for impulsive reconstruction is a pitch detection algorithm for NB excitation, which makes use of the autocorrelation to find the pitch frequency. Once the pitch frequency is determined, impulses are placed at pitch harmonics in the frequency domain as seen in Fig.2. However, this is only done for the low frequencies, 0–300Hz, as for 3400-4000Hz the frequency content can be assumed to be spectrally flat, thus leading to extension via simple modulation.

In Fig. 3, an impulsive reconstruction is performed on the topmost plot to obtain the plot in the middle, which approximates the original WB excitation. As also seen in the figure, most of the important information is carried in the low frequency band. The magnitudes of the impulses that are being placed at pitch harmonics are decided with a moving average filter performed on the magnitude spectrum of NB excitation within frequencies 300Hz-1000Hz. For the high frequency content, however, the simple modulation technique is used. Once every frame's excitation signal is estimated separately, the whole sentence is reconstructed by convolving the excita-

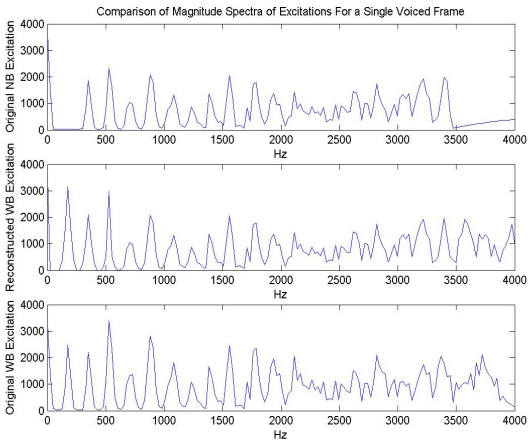

Fig. 3. Comparisons of Magnitude Spectra for Different Excitation Signals of a Single Frame

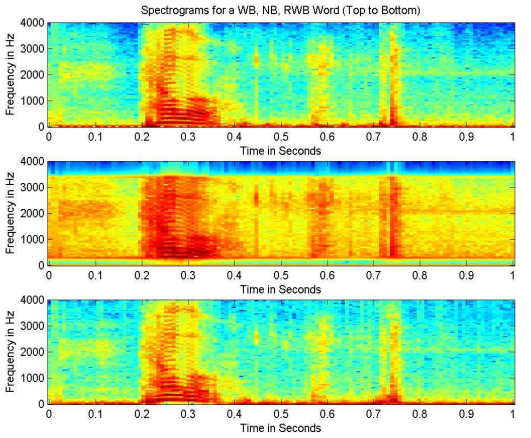

Fig. 4. Comparisons of Spectrograms for the word "Hello"

tions with the respective spectral envelopes for each frame. Fig. 4 shows the original WB, original NB and RWB versions of the word "Hello." In Fig. 4, as compared to NB speech, RWB speech has more relevant low and high frequency content. In terms of perceptual sound quality, RWB sounds closer to WB speech in terms of naturalness, as compared with NB speech. See section 4 for more details.

3. EXTENSION OF SPECTRAL ENVELOPE

Similar to the excitation extension, in order to isolate the spectral envelope extension, this algorithm utilizes the WB excitation and extrapolates the NB spectral envelope to that of the RWB spectral envelope. The spectral envelope extension problem is basically a problem of finding the right feature set and the right mapping technique between NB and WB feature sets. These two components are discussed in the following sections.

3.1 Feature Extraction Techniques

The first step is to find the appropriate feature extraction technique to create the feature set. The Linear Source Filter Model (LSFM) is an all-pole filter model with coefficients corresponding to the Linear Prediction Coder (LPC) coeffic-

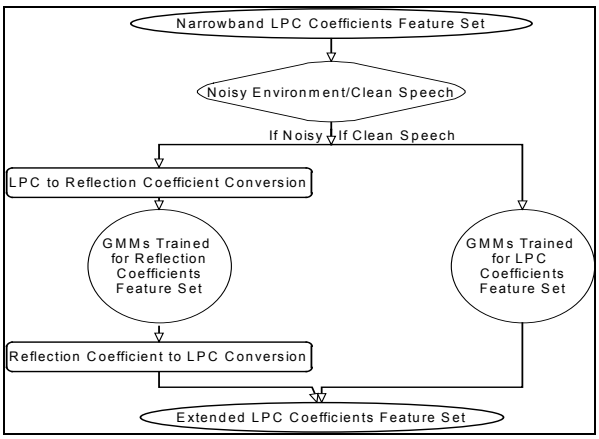

Fig. 5. Block Diagram of Spectral Env. Extension Algorithm

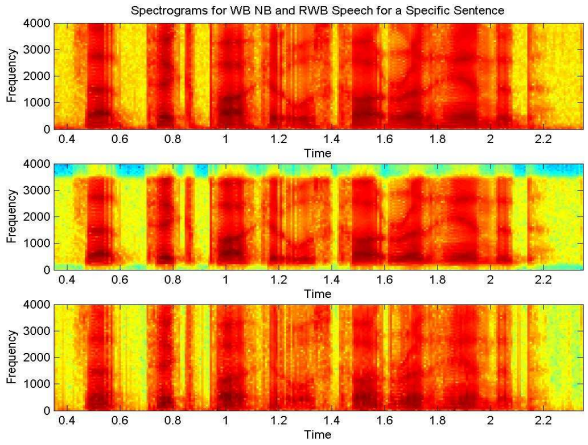

Fig. 6. Comparisons of Spectrograms for a Sentence

cients [1]. LPC feature extraction is crucial when using the LSFM, and constitutes a viable technique of extraction. However, LPC coefficients are extremely sensitive to noise. More robust feature extraction methods include poles of the spectral envelope, log areas, and reflection coefficients. These feature sets also are beneficial because of their one-to-one mapping with LPC coefficients.

The poles of the vocal tract are theoretically a more robust space than just LPC. However, after the mapping between the NB and WB feature sets, some poles, which are close to the unit circle fall just outside. This scenario constitutes an unstable system and thus the poles are not a viable feature set. The log area coefficients represent the cross sectional areas of the vocal tract segments [4]. However, experimental results show that, overall, the log area coefficients are worse than LPC and are also not used as a feature set.

The final intermediate feature space considered is reflection coefficients (RCs) [4]. RCs give the relative amplitudes of the incident and reflected pressure waves at the juncture between the tube segments. When compared to Log Areas, RCs provide better speech output. Also, in the presence of noise, the use of RC intermediates results in better quality speech than just the use of LPC coefficients.

3.2 Mapping Techniques

In speech applications, Hidden Markov Models (HMMs) are used widely for non-stationary classification and mapping purposes [5]. However, the classification window in this algorithm creates a situation where speech frames are stationary and thus no more than one state is needed in modelling a frame. A single state HMM with an output probability density function of a jointly Gaussian random variable can be equivalently modelled as a Gaussian Mixture Model (GMM) [6]. Overall, GMMs are also less computationally complex than HMMs. Consequently, the mapping technique used is GMMs.

A GMM is based on a collection of input data (i.e. LPC frame features) and training this data with a specified number of component Gaussians. An iterative algorithm is used to refine the GMM parameters, such that with subsequent iterations, there is a monotonic increase in the likelihood parameters given a feature set. Hence, for a given training set of feature vectors, the maximum likelihood parameters are estimated iteratively using the Expectation Maximization (EM) algorithm [7].

For a D-dimensional feature vector (i.e. LPC) x, the mixture density used for the likelihood function is:

$$p(x/\lambda) = \sum_{i=1}^{M} w_i p_i(x)$$

The density, $p_i(x)$ is a weighted linear combination of M Gaussian densities with mean vector μ and covariance matrix Σ_i.

$$P_i(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'(\sum_i)^{-1}(x-\mu_i)\right\}$$

The mixture weights are constrained to $\sum_{i=1}^{M} w_i = 1$. The parameters of the density model are $\lambda=\{w_i, \mu_i, \Sigma i\}$,where i = 1,….M

As also shown in Fig.5, if needed, the WB features are converted back to LPC, if LPC intermediates are used as the feature extraction technique. This Robust Spectral Envelope Extension (R-SEE) algorithm uses the convolution between these extended LPC features, as well as the WB excitation to attain RWB speech.

Fig. 6 shows from the top to bottom, the spectrograms for the WB, NB, and RWB speech sentence "That pickpocket was caught red handed." The RWB speech is found through the convolution of WB excitation with the estimated extended spectral envelope, which is based upon GMMs that are trained using the EM algorithm. The feature extraction set is LPC coefficients in a noise-free environment. Note, there is a lack of frequency information between the bands of 0-300 Hz and 3400-4000 Hz in the NB speech sentence. The algorithm extends the NB speech properties into these bands, such that relevant frequency information is added, as seen by the similarity between the RWB and WB spectrograms.

4. QUANTITATIVE EXPERIMENTAL RESULTS

4.1 Performance Metrics

Even though there is no universal objective measure to determine the quality of speech, in order to estimate the accuracy of any BWE algorithm, the RMS log spectral measure can be used [8].

$$D^2_{WB->NB} = \frac{1}{N}\sum_{k=1}^{N}\left\{10\log_{10}\left|X_{WB}(k)\right| - 10\log_{10}\left|X_{NB}(k)\right|\right\}$$

where N is the total number of samples and $\left|X_{WB}(k)\right|$ denotes the absolute FFT of the speech frame. The distortion metric is then averaged for all frames in the sentence.

Sent #	$D_{WB->NB}$	$D_{WB->RWB(\mathrm{Prev.})}$	$D_{WB->RWB(F-BEE)}$
1	9.29	5.44	4.57
2	11.45	5.87	4.25
3	11.39	7.55	5.38
4	8.13	6.60	4.91
5	12.89	5.47	3.91
6	9.12	5.18	4.85
Avg:	10.38	6.02	4.64

Table. 1. Distortion Metric Results for Excitation

Table.1 shows the difference in distortions first when no excitation extension algorithm is present, second when the excitation modulation technique is used, and finally when the F-BEE is used. The six sentences are recorded in MATLAB™ with f_s = 16kHz. The rightmost column has the lowest distortion, which gives rise to better speech.

Sent #	$D_{WB->NB}$	$D_{WB->RWB(LPC)}$	$D_{WB->RWB(RC)}$
023	10.39	3.86	4.20
067	10.37	3.88	4.62
136	10.69	4.87	4.35
172	10.46	4.29	5.00
208	10.46	3.98	4.99
291	10.61	4.28	6.69
318	10.42	4.10	5.18
377	10.11	3.40	4.05
Avg:	10.44	4.08	4.89

Table. 2. Distortion Metric Results for Spectral Envelope Extension

Table 2 uses the same distortion metric and shows the difference in distortions first when no spectral envelope excitation is used, second when LPC features are used, and finally when RC intermediates are used. The noise-free sentences are from the TIMIT™. The rightmost column shows that even in noise free environments RC features perform comparably well with LPC features.

BWE type:	$D_{WB->RWB}$
Vector Quantization (VQ) [9]	As low as 10.6
Previous GMM algorithms	As low as 7.4
R-SEE algorithm	3.4-4.9

Table. 3. Comparison of Distortion Metrics for Different Alg.

The distortion metric between 300-34000Hz and 0-4000Hz speech is greater than that of 0-4000Hz and 0-8000Hz speech as a result of the inherent property of speech, which states that most of the frequency information is stored in lower frequencies. This indicates that the extension of 300-3400Hz to 0-4000Hz speech is more difficult than the extension of 0-4000Hz to 0-8000Hz speech. Even though this extension problem is more complicated, the results shown in Table 3 indicate that the R-SEE algorithm still manages to outperform other spectral envelope extension algorithms.

5. CONCLUSIONS AND FUTURE WORK

This paper introduces a novel approach to the bandwidth extension problem. The algorithm utilizes the LSFM approach, which breaks down speech into two components, the excitation and the spectral envelope. Both components are extended separately via the F-BEE and the R-SEE algorithms developed in this paper to estimate RWB speech, which has more naturalness than the NB speech. Experimental results also show that, each extension algorithm has less distortion by themselves than previous extension algorithms. As future work, the incorporation of these two algorithms will be performed. Further noise analysis will be done to stress the advantages of using reflection coefficients under noisy conditions.

6. ACKNOWLEDGMENTS

This research project is funded by Motorola™, Inc., and has been carried out at the Computational NeuroEngineering Lab (CNEL) of the University of Florida, Gainesville, USA.

REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637-655, Aug. 1971.

[2] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," *Proc. EUSIPCO 1994*, Edinburgh, Scotland, U.K., September 13-16, pp. 1178-1181.

[3] Jr. A. Gray and J. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, pp. 207-217, June 1974.

[4] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs: Prentice Hall, 1973.

[5] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden Markov model," *IEEE Workshop on Speech Coding*, Delavan, Wisconsin, Sept. 17-20. 2000, pp. 133-135.

[6] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," *ICASSP 2004*, Montreal, Quebec, Canada, May 17-21, pp. 697-700.

[7] A. Dempster, N. Laird and D. Rubin, "Maximum likelyhood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol.39, pp. 1-38, 1977.

[8] Jr. A. Gray and J. Markel, "Distance Measures for Speech Processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380-391, Oct. 1976

[9] A. Buzo, Jr. A. Gray and J. Markel, "Speech coding upon vector quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 562-574, Oct. 1980.