SPEECH BANDWIDTH EXTENSION: EXTRAPOLATIONS OF SPECTRAL ENVELOP AND HARMONICITY QUALITY OF EXCITATION

Saeed Vaseghi, Esfandiar Zavarehei, Qin Yan

School of Design and Engineering, Brunel University, London UK
{saeed.vaseghi, esfandiar.zavarehei, qin.yan}@brunel.ac.uk

ABSTRACT

This paper presents a method for restoration of the missing bandwidth of narrowband speech signals. Speech is decomposed into a linear prediction (LP) model of the spectral envelop and a harmonic plus noise model (HNM) of speech excitation. The LP spectral envelope and HNM excitation parameters of the narrowband speech are extrapolated using codebooks trained on narrowband and wideband speech. A novel contribution of this paper is the introduction of a parametric measure of the *harmonicity* of excitation in harmonically-spaced sub-bands. The wideband LSF parameters and the degree of harmonicity of missing excitation are estimated from those of the narrowband speech via codebook mapping. The method is successful in restoring the harmonicity of speech and converts telephone quality speech to perceptually high quality wideband speech.

1. INTRODUCTION

The extrapolation and interpolation of missing spectrum of audio and speech signals has applications in speech telecommunication and restoration of band-limited archived speech recordings. Telephony speech is limited to a bandwidth of less than 4 kHz; normally about 3.4 kHz. In recent years several methods have been proposed for bandwidth extension of telephony speech signals to a higher bandwidth of broadcast quality. The aim of these methods is to reconstruct the upper band contents of speech signals using an extrapolation of the spectral envelop from the available spectrum at lower bands in order to gain the sensation of higher bandwidth and higher quality speech.

Most bandwidth extension techniques strive to reproduce the expanded spectral envelop through the use of codebook mapping methods [1-3]. The missing spectral envelope in the higher bands is obtained from codebooks trained on joint feature vectors of band-limited and full band speech. The spectral envelop representation for codebook mapping is often based on the line spectral frequencies (LSF) parameters derived from a linear prediction model of speech [3]. The estimates of the spectral envelop are then combined with an estimate of the excitation signal to yield the wideband output speech signal.

Different methods are used for estimation of the excitation signal, such as spectral folding, Gaussian modulation, etc [1][4]. The main difficult challenge for these algorithms is to recover those parts of signal in which valuable information resides in the upper band rather than the lower band (e.g. fricatives).

In this paper a new approach is proposed for estimation of the excitation for the missing parts of speech spectrum. For voiced speech the excitation harmonics are tracked and extrapolated. An essential aspect of the reproduction of the harmonics is the idea of the extrapolation of the quality of harmonics of speech [1]. For this purpose a measure of harmonicity of speech excitation is defined. This measure is used to characterize both the harmonic structure of voiced speech and also the non-harmonic structure of unvoiced speech. This approach has the added advantage that it circumvents the need for hard decisions regarding the classification of speech into voiced/unvoiced segments/sub-bands.

Figure 1 illustrates the overall block diagram of the bandwidth extension system. The excitation and the LP model of each frame are extended using codebook mapping. The wideband speech signal is reconstructed by combining the excitation and the LP model of the envelope of the signal.

The rest of the paper is organized as follows: In section 2 the linear prediction - harmonic plus noise model (LP-HNM) of speech is described. Section 3 describes modeling and estimation of the excitation and introduces the use of the harmonicity measure for this purpose. The estimation of the LP model of the spectral envelope of the signal and its gain value is discussed in section 4. In section 5 the experimental results are discussed. Conclusions are drawn in section 6.

2. LP-HNM MODEL OF SPEECH

The model used for speech bandwidth extension is a linear prediction (LP) model of spectral envelop and a harmonic noise model (HNM) of speech excitation. In frequency domain the LP-HNM model of the speech magnitude spectrum may be expressed as:

$$X(f) = G \cdot E(f) / A(f) \tag{1}$$

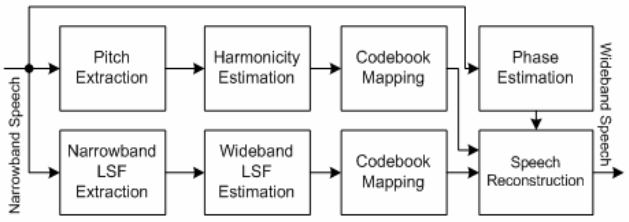

Figure 1. Block diagram of the bandwidth extension system

where $E(f)$ is the excitation signal, $1/A(f)$ is a LP model of the combined effect of the vocal tract, the glottal pulse and the lip radiation and G is a gain factor. The excitation can be modeled by a mixture of harmonics and noise as

$$E(f) = \sum_{k=1}^{Nharmonics} A_k M\left(f - kF_0\right) + N(f) \tag{2}$$

where F_0 is the fundamental frequency of the excitation, $M(f)$ models the shape of each excitation harmonic which may be set to a delta function or more realistically to a Gaussian-shaped waveform and $N(f)$ is the noise component of the excitation.

3. A HARMONICITY MODEL OF EXCITATION

The excitation signal is modeled using an HNM model as the summation of harmonics and noise components. Rather than using a hard decision for the classification of the sub-bands to voice/unvoiced, the ratio of the harmonic energy to the noise energy in each sub-band is calculated as the level of *harmonicity* of that sub-band:

$$H_n = 1 - \frac{\int_{nf_0 - f_0/2}^{nf_0 + f_0/2} \left[\, \left|E\left(nf_0\right)\right| M\left(f - nf_0\right) - \left|E\left(f\right)\right| \,\right]^2 df}{\int_{nf_0 - f_0/2}^{nf_0 + f_0/2} \left|E\left(f\right)\right|^2 df} \tag{3}$$

where H_n is the harmonicity of the excitation signal in the n^{th} band, $E(f)$ is the DFT of the excitation signal and $M(f)$ is an arbitrary function that models the shape of harmonics. Kondoz [5] proposed the Fourier transform of hamming window to be used as the harmonic function $M(f)$. In this work a Gaussian shaped window is used which except for some scaling is equal to its Fourier transform. This function is shown in Figure 2.

$$M\left(f\right) = \exp\left(-\left(f/45.45\right)^2\right) \tag{4}$$

The use of the degree of harmonicity for this application, results in a soft decision during codebook mapping rather than a hard decision. Moreover, during temporal fluctuations of the spectrogram, the harmonics gradually gain or lose strength which is modeled using the harmonicity degree. The excitation of each frame then is reconstructed as:

$$\left|\hat{E}\left(f\right)\right| = E\left(nf_0\right)\left(\frac{H_n M\left(f - nf_0\right)}{\sqrt{\int M^2\left(f\right) df}} + \frac{\left(1 - H_n\right) N\left(f\right)}{\sqrt{\int N^2\left(f\right) df}}\right) \tag{5}$$

$$\text{for} \qquad nf_0 - \frac{f_0}{2} < f < nf_0 + \frac{f_0}{2}$$

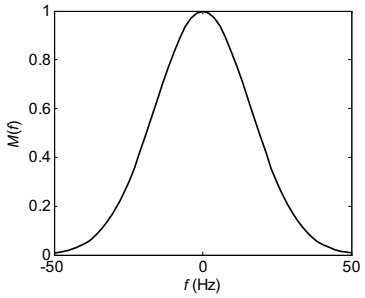

Figure 2. A Gaussian $M(f)$ was used for modeling harmonics

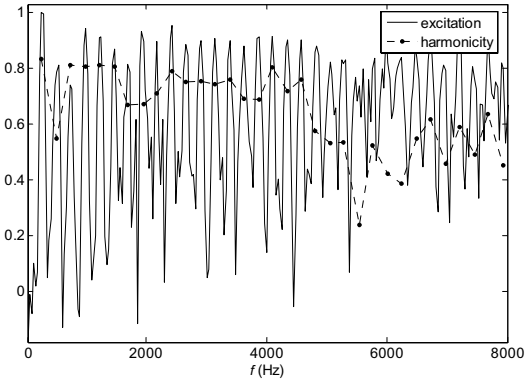

Figure 3. Harmonicity of the excitation sub-bands superimposed on the normalized excitation

where $N(f)$ is a Rayleigh distributed random variable to comply with the assumption of the Gaussian distribution model of the speech DFT [6]. Figure 3 illustrates the excitation of a sample frame together with the harmonicity values of each band. The spectrum of this frame is illustrated in Figure 4 (top) and the reconstructed spectrum is depicted in Figure 4, (bottom). During the reconstruction of the high frequency part of the spectrum, the excitation amplitudes in each sub-band, at the harmonics $E(nf_0)$, is assumed to be 1 which is reasonable assuming that the order of the LP model is high enough to have whitened the excitation. The gain of the frame is modeled using the LP gain value. The excitation reconstructed from this model is multiplied by the LP spectrum and the phase component to form the speech spectrum of each frame. For the codebook to have a fixed size, the harmonicity of a maximum of 32 bands are calculated for each frame. Sub-bands above this range are assumed to be entirely unvoiced. The efficiency of the system seems to saturate for larger number of harmonics. A codebook of harmonicity feature vectors is trained on 16 kHz speech signals from Wall Street Journal (WSJ) speech data base using the K-means algorithm. The distortion measure is the Euclidian distance measure which is applied only to the lower 4 kHz of the signal. The upper 4 kHz harmonicity values act as a shadow codebook for retrieving the harmonicity of the upper sub-bands.

Spectral matching method is used for fundamental frequency extraction which is essential for harmonicity calculations [5]. An

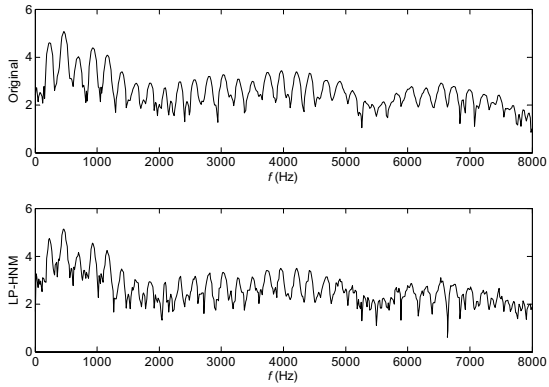

Figure 4. Harmonicity based reconstruction: Original (top) and LP-HNM Reconstructed (bottom)

Table 1. Average log spectral distance of the bandwidth extended signal from the wideband signal for energy normalization and codebook mapping of the gain factor.

Distortion Measure (dB)	Energy Normalization	Codebook Mapping
Overall FFT-LSD	5.83	5.10
High band FFT-LSD	8.25	7.17
Overall LP-LSD	4.96	4.27
High band LP-LSD	6.83	6.09

initial fundamental frequency is first calculated using spectral matching, which is then refined by pick peaking and finding the harmonic frequencies and adjusting the fundamental frequency according to harmonic frequencies. Furthermore, Viterbi algorithm is used to *track* the fundamental frequency and pick the best candidate among different candidates [7].

4. LP ENVELOPE EXTRAPOLATION

The LP model coefficients are converted to line spectral frequency (LSF) coefficients as these have good quantization and extrapolation qualities. A 12th order speaker-independent LSF codebook is trained based on narrowband speech signals derived from wall street journal (WSJ) speech database. A shadow codebook of 24th order 16 kHz LSF is trained in conjunction with the 8 kHz codebook. The distortion measure used both in training stage and in estimation stage is the mean square spectral distortion (error) of the corresponding LP spectra:

$$D(X_{LSF}, Y_{LSF}) = \int_{-0.5}^{0.5} \left| X_{LP}(f) - Y_{LP}(f) \right|^2 df \qquad (6)$$

where X_{LSF} and Y_{LSF} are the set of narrowband LSF vectors and $X_{LP}(f)$ and $Y_{LP}(f)$ are the corresponding LP spectra.

Estimation of the gain of the LP model, G, is crucial in bandwidth extension. Two different approaches for gain estimation were implemented: i) estimation of the gain using energy normalization and ii) codebook mapping. In the first approach the gain of the wideband LP model is calculated to result in the same amount of energy in the low-band portion as that of the narrow band speech signal. While this approach works well in the voiced segments, experiments show that it does not result in good estimates for the frames where most of the energy is concentrated in the high frequency (e.g. fricatives).

In the second approach, the ratio of the gain of the narrowband LP model of each frame, G_8, divided by that of the wideband signal for the same frame, G_{16}, is calculated:

$$R_G = G_8 / G_{16} \qquad (7)$$

The gain ratio R_G is calculated for each frame and a shadow codebook is trained in conjunction with the LSF codebook, similar to the shadow codebook for the wideband LSF.

During the estimation stage, the narrowband LSF values are calculated for each frame and the closest (measured in terms of spectral distortion) codeword is chosen from the narrowband LSF codebook. The corresponding wideband codeword and gain ratio are obtained from the shadow codebooks. The narrowband LP model gain is divided by the gain ratio and the result is used as the wideband LP model gain.

Table 2. Average log spectral distance of the bandwidth extended signal from the wideband signal for band-pass envelope modulated Gaussian noise (BP-MGN) model and Harmonicity model.

Distortion Measure (dB)	BP-MGN Model	Harmonicity Model
Overall FFT-LSD	5.85	5.10
High band FFT-LSD	8.79	7.17
Overall LP-LSD	4.51	4.27
High band LP-LSD	6.63	6.09

5. EXPERIMENTS

The LSF, gain and harmonicity codebooks are trained using speech signals obtained from WSJ speech database, spoken by several speakers. The WSJ speech is originally sampled at 16 kHz and has a bandwidth of 8 kHz. Speech is segmented to frames of 25 ms duration with a frame overlap of 15 ms i.e. a frame rate of 10 ms. To produce narrowband speech, wideband speech is filtered to a bandwidth of 4 kHz and down-sampled to a sampling rate of 8 kHz.

For the purpose of the evaluation, 40 test sentences are randomly chosen from the WSJ speech database, the test sentences were not among those used for training the codebooks. The mean log spectral distance (LSD) between the wideband and bandwidth-extended signal are calculated and averaged over all frames and sentences for each modification. The average LSD is calculated using the FFT of the frames (FFT-LSD) and LP spectrum of the frames (LP-LSD).

There are several different methods for coding and estimation/prediction of the phase of the missing spectra such as phase codebooks and phase predictors [8]. As most of the signal in the higher bands of the wideband speech is not harmonically structured, the issue of phase estimation is not extensively explored in the literature of bandwidth extension. In this work the phase of the upper band is estimated from the lower band so that the unwrapped phase of each frame is linear. Some random phase, proportional to the inverse of the harmonicity of each sub-band is then added to the phase to account for the non-harmonic random phase [9]. In experimental evaluations, absolutely no perceptible difference is audible when the upper band phase of a wideband signal is replaced by its predicted value.

5.1. Codebook Mapping Of the Gain

The estimation of the LP gain of the wideband signal is crucial in bandwidth extension of narrowband signals. The use of codebook mapping for estimation of the LP gain is compared with energy normalization method. It is observed that using energy normalization results in suppression of the signals, especially during fricative segments where most of the energy of the signal is concentrated in the higher bands of the signal. Estimation of the LP gain value through codebook mapping results in superior quality of the wideband signals. A comparison of the LSD values of these two methods is presented in Table 1 which shows that codebook mapping results in lower averaged LSD distances in every case.

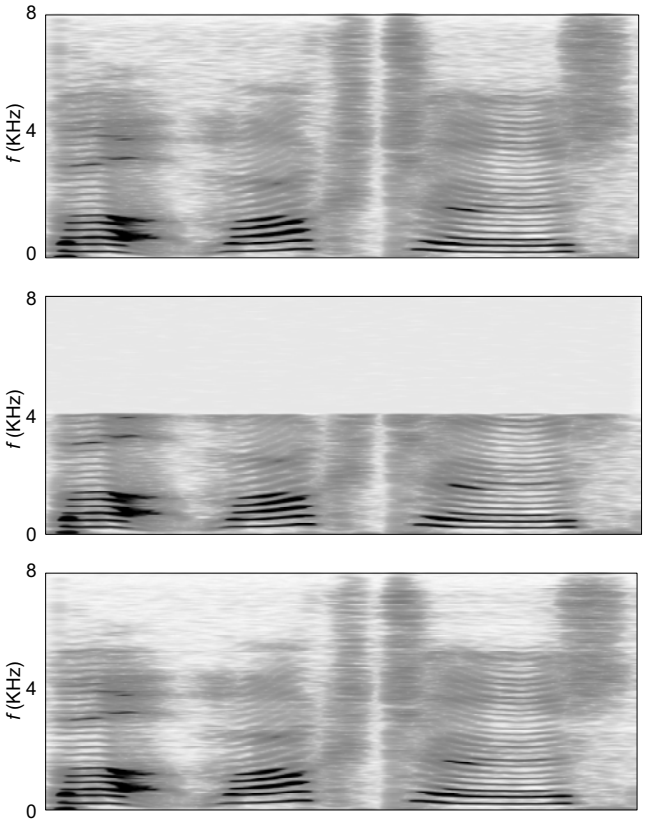

Figure 5. Spectrograms of (top) wideband signal (middle) Narrowband Signal and (bottom) bandwidth extended signal

5.2. Harmonicity Model

The use of the harmonicity model for reconstruction of the excitation of the signal is compared to the band pass envelope modulated Gaussian noise (BP-MGN) method used in [1] and [3]. While the rest of the system is similar for evaluations carried out here, only the excitation is estimated using the two different methods. It is observed that both methods result in reasonably good quality output speech. However, from the study of the spectrograms of the wideband signals produced by these systems it was observed that the extended bands which had more harmonically structured patterns were better reconstructed using the proposed harmonicity model. Table 2 summarizes the averaged LSD values calculated for these two cases. While the difference linear prediction based LSD's are large, those of the FFT-LSD seem to be higher. We believe that this is due to the more detailed modeling of the harmonics of the higher bands in the proposed method.

Figure 5 shows an example of the original wideband speech with a bandwidth of 8 kHz, the filtered speech with a bandwidth of 4 kHz and the restored extended bandwidth speech. LP Gain and harmonicity values are estimated using the proposed method of codebook mapping. The figure clearly shows both the harmonic and noise structures in the upper band have been well restored.

5.3. Sensitivity to Pitch

To accurately estimate the harmonicity levels, it is crucial that reasonably accurate estimates of the fundamental frequency are extracted from the narrowband speech signal [10]. Inaccurate fundamental frequency estimates normally do not happen in harmonically well-structured speech frames, which are also more likely to have higher frequency harmonics. It is observed that inaccurate fundamental frequencies during harmonically structured (voiced) frames results in inaccurate extrapolation of harmonicity. However, this is not the case during unvoiced frames. During unvoiced frames the excitation signal is normally reconstructed using only noise as even if a fundamental frequency is assigned to it the harmonicity values calculated for sub-bands will be very low which will result in noise domination in Equation (5).

6. CONCLUSION

A method for bandwidth extension of narrowband signals was introduced. The effect of incorporation of the harmonicity measure, and extrapolating these values for estimation of the excitation signal was investigated and compared with band pass modulated Gaussian noise method used for wideband excitation estimation. The LP-HNM model has the ability to revive the missing harmonics of speech that reside in the upper frequency band. Its potential for reviving those harmonics in the lower frequency band is being investigated. The immediate application of this would be speech enhancement and also restoration of the old archived recordings.

7. REFERENCES

[1] Qian and Kabal, "*Dual-Mode Wideband Speech Recovery from Narrowband Speech*", Proc. Eurospeech 2003, pp. 1433-1436.

[2] D. G. Raza and C.-F. Chan, "*Enhancing Quality of CELP Coded Speech via Wideband Extension by Using Voicing GMM Interpolation and HNM Re-Synthesis*", Proc.Int. Conf. Acoustics Speech and Signal Processing, pp. I-241–I-244, 2002.

[3] R. Hu, V Krishnan, D. V. Anderson, "*Speech bandwidth extension by improved codebook mapping towards increased phonetic classification*", Interspeech 2005 – pp. 1501-1504

[4] J. P. Cabral, L. C. Oliveira, "*Pitch-Synchronous Time-Scaling for High-Frequency Excitation Regeneration*", Interspeech 2005 – pp. 1513-1516

[5] A. M. Kondoz, "*Digital Speech: Coding for Low Bit Rate Communication Systems*", 2nd Edition, John Wiley & Sons, 1994

[6] Ephraim, Y., Malah, D., "*Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator*", IEEE Trans. ASSP on Acoustics, Speech, and Signal Processing, vol. -32, no. 6, pp. 1109-1121, Dec. 1984.

[7] H. Quast, O. Schreiner, M.R. Schroeder, "*Robust Pitch Tracking in the Car Environment*", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002 Orlando, Vol 1 1pp. I-353- I-356

[8] Hui YE, "*Voice Morphing Using a Sinusoidal Model and Linear Transformation*", PhD Thesis, Cambridge University Engineering Department, May 2005.

[9] DVSI. *INMARSAT M VOICE CODEC*. USA February 1991, Version 1.3.

[10] D.W. Griffin, J.S. Lim, "*Multiband-excitation vocoder*", IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-36:2, pp. 236--243, 1988.