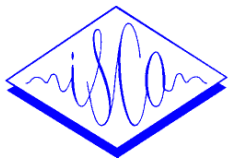

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

BEYOND NYQUIST: TOWARDS THE RECOVERY OF BROAD-BANDWIDTH SPEECH FROM NARROW-BANDWIDTH SPEECH

Carlos Avendano, Hynek Hermansky, and Eric A. Wan

Oregon Graduate Institute of Science & Technology
Department of Electrical Engineering and Applied Physics
P.O. Box 91000, Portland, OR 97291

ABSTRACT

A new technique is presented which improves the subjective quality of band-limited speech. The approach is based on a linear model of speech production, in which we independently estimate the spectral envelope and excitation function for a broad-bandwidth speech signal to reconstruct missing frequency components in narrow-bandwidth speech.

10.21437/Eurospeech.1995-44

1. INTRODUCTION

To save on frequency spectrum in public switched telephony, only the minimum amount of signal bandwidth necessary to preserve intelligibility of speech is transmitted. Recovering missing frequency components at the receiver improves the subjective speech quality and is of a significant commercial interest. Typically, telephone speech is limited to frequency components in the 300-3300Hz range. In the ideal case we would like to recover the components between 3.3kHz and 8kHz, as well as the components in the low 0-300Hz frequency band from a sample of telephone speech.

Nyquist theory states that the sampling rate places a fundamental limit on the frequencies that can be recovered for arbitrary signals. However, speech is produced by a well defined physical system. Consequently, the higher and lower spectral components which are missing in the telephone speech have a direct relation to the spectral components already present. We currently do not understand this relation well enough to offer an analytic solution to the recovery problem. However, we previously demonstrated that by taking advantage of the correlation between low and high frequency components in the short-term power spectrum of speech from a given speaker, high frequency recovery appears to be viable [1]. In that study, an attempt was made to predict each frequency component in the 4-8kHz range by mapping time trajectories of the cubic-root compressed short-term power spectrum of a 4kHz band-limited speech signal. In this work we attempt to predict only the spectral envelope of the short-term spectrum and derive the spectral fine structure from the available speech.

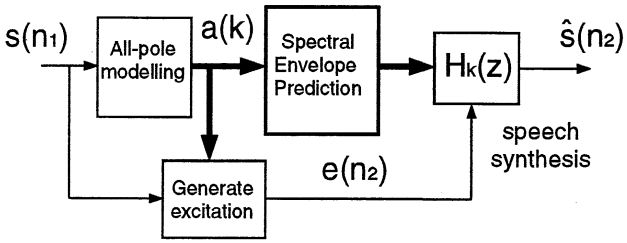

Figure 1: *Basic building block of the system*

2. APPROACH

With the current technique we attempt to recover high frequency components up to 8kHz, from a 4kHz low-pass filtered speech signal. We also try to recover the low frequency band from 300-3300Hz band-pass filtered speech. In the following discussion, we will first describe the technique used to recover the high frequency band, and later we will point out the changes in the system needed to accomplish the low band reconstruction. The principle under which our technique works is to produce a synthetic signal derived from the narrow-band speech, containing an estimate of the missing frequency components.

This current work is based upon the linear model of speech production, i.e. the synthetic signal is produced by independently estimating its spectral envelope and its spectral fine structure. In this way, the dimensionality of the mapping is significantly reduced compared to earlier techniques [1],[2]. Additionally, the excitation signals for generating the fine spectral structure could be derived using techniques adopted from current residual-excited and stochastic codebook-excited speech coding. The synthetic speech signal is properly

mixed with the narrow-band speech sample to produce the reconstructed speech.

3. HIGH FREQUENCY RECOVERY

The broad-band spectral envelope is predicted from a time window of the spectral envelope of the available narrow-band speech. This broad-band envelope covers 0-8kHz and is used to estimate the frequency components in the 4-8kHz high frequency band. With the given predicted envelope, we proceed to design an all-pole synthesis filter. The excitation function for this filter is derived from the narrow-band speech sample.

Short-term analysis of a narrow-band speech signal $s(n_1)$ is done by taking 25ms time windows with a 10ms frame rate (n_1 corresponds to 8kHz sampling rate). An all-pole model $[\sum_{i=0}^{p} a_i(k)z^{-i}]^{-1}$ is derived for each frame k,using the autocorrelation method. We are currently using an 8th order model for the 8kHz sampled signal. This model is then used for the spectral envelope prediction, and in the process of estimating the excitation signal.

As illustrated in figure 1, an all-pole synthesis filter $H_k(z)$ is derived from the predicted spectral envelope and is excited by a proper excitation $e(n_2)$ (n_2 corresponds to 16kHz sampling rate). During synthesis, energy matching between the original signal and the synthetic signal is accomplished by scaling the excitation in a similar way as described in [3]. The resulting synthetic signal $\hat{s}(n_2)$ is filtered and mixed properly with $s(n_1)$ to produce the recovered broad-band speech.

3.1. Envelope Prediction

The envelope prediction is done by filtering time trajectories of LPC-cepstral coefficients of the narrow-band signal $s(n_1)$ through a multidimensional filter designed on training data. The short-term power spectrum of speech is modeled by an all-pole system at each analysis frame. During training this can be performed for the narrow-band speech $s(n_1)$ and the desired broad-band speech $s_d(n_2)$. For $s_d(n_2)$ we generally use a higher order model (currently 16th order) than the order of $a(k)$.

Cepstral coefficients for the two sets of autoregressive coefficients are computed and the time trajectories of these coefficients are then used to design a bank of multi-input single-output filters. Each filter of the bank is designed to map a time window of the 8 coefficient trajectories from $s(n_1)$ to a particular coefficient corresponding to the current frame of $s_d(n_2)$. Thus, N (N=16) such filters constitute the filter bank.

Let $C_i(k)$ (i=1,2,..,p) be the ith cepstral coefficient of $a(k)$. The output of each filter is the following:

$$\hat{C}_r^d(k) = \sum_{i=1}^{p} \sum_{l=-M}^{M} W_{i,r}(l) C_i(k-l), \qquad (1)$$

where $\hat{C}_r^d(k)$ is the estimate of the rth cepstral coefficient corresponding to the envelope of $s_d(n_2)$ (r = 1,2,..,N and k corresponds to a 10 ms step). FIR filter coefficients $W_{i,r}$ are found such that $\hat{C}_r^d$ is the least squares estimate of the original C_r^d. The current design uses a 100ms time window (*i.e.* $M = 5$).

3.2. Excitation Signal

Using available information about the fine spectral structure in $s(n_1)$, we can derive an excitation function $e(n_2)$ to drive the synthesis filter $H_k(z)$.

The excitation $e(n_2)$ for the high frequency band reconstruction is generated by up-sampling by two and whitening the residual signal $r(n_1)$ at each frame [4]. The residual results from filtering $s(n_1)$ with the inverse filter whose impulse response is given by $a(k)$. The spectral folding operation on $r(n_1)$ does not generate a signal with a regular harmonic structure during voiced frames because the fold occurs at the Nyquist frequency regardless of the last harmonic frequency position, however no adverse effects were heard. Frames with high energy in the high frequency band are generally fricatives and the fine structure has little or no harmonic structure at all.

We tried other techniques for higher harmonics regeneration (see also [4]), but the method described above has given the best results to date.

3.3. Recovered broad-band speech synthesis

The synthetic signal produced by driving the synthesis filter $H_k(z)$ with the excitation $e(n_2)$ is mixed with $s(n_1)$ in the following way: The narrow-band signal $s(n_1)$ is up-sampled by two and filtered by the low pass section of a quadrature mirror filter. The synthetic signal $\hat{s}(n_2)$ is filtered by the high-pass section of the mirror. Both filtered signals are added and the result is a full bandwidth (0-8kHz) signal at a 16kHz sampling rate.

4. LOW FREQUENCY RECOVERY

For the envelope prediction, we train the multidimensional filter in the same way described above. The difference is the order of the all-pole model used. Now $s_d(n_2)$ is an 8kHz sampled speech signal ($n_2 = n_1$) and

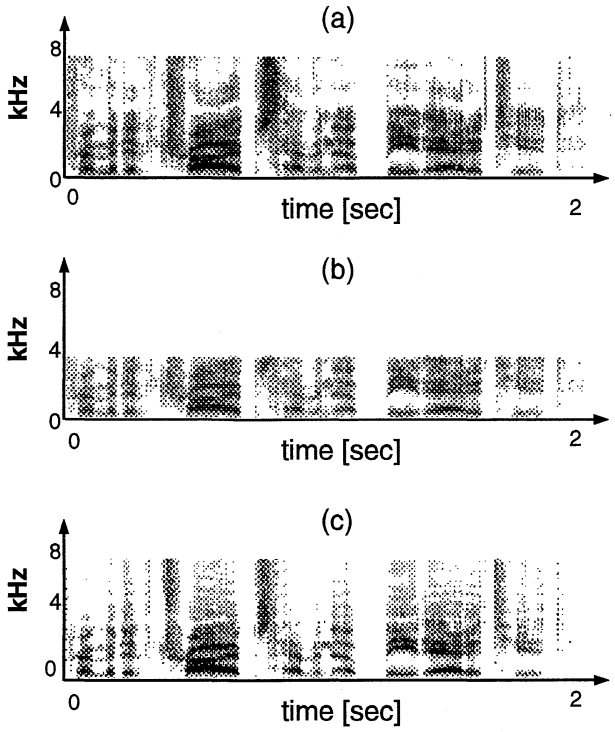

Figure 2: *Spectrograms showing (a) original speech followed by (b) the narrow-band version and (c) reconstructed segment after processing.*

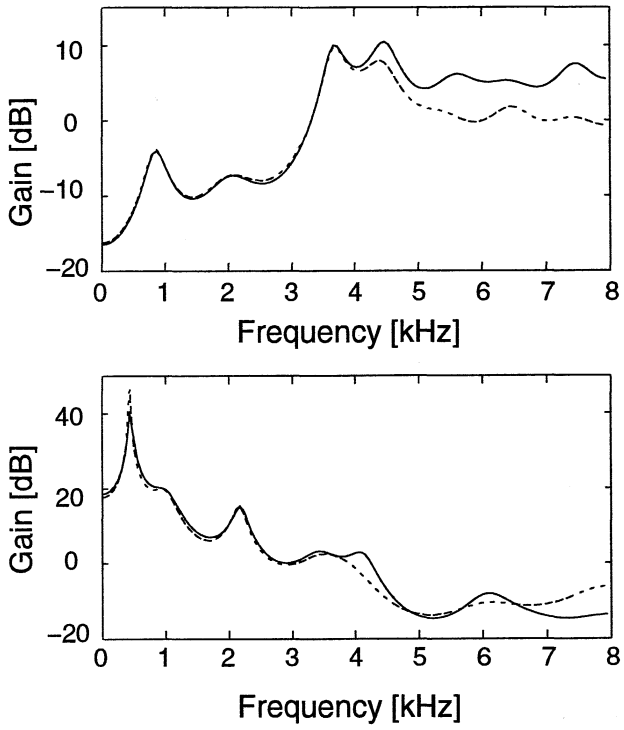

Figure 3: *Original (dark line) vs. predicted (dashed line) envelopes for the high frequency band.*

$s(n_1)$ is a band-pass filtered (300-3300Hz, telephone-like) signal. Both models used are of the 8th order (p=N=8 in (1)).

For the low frequency band excitation, the problem lies in trying to regenerate the fundamental frequency (and possibly the first few harmonics) in the residual $r(n_1)$ of voiced segments. We tried several methods to solve this problem. As it is often done in telephone speech pitch tracking algorithms, a non-linear distortion can be applied to the residual. Although this technique is successful in re-introducing the lower frequencies, the resulting synthetic speech has a harsh sound. Finding the pitch (a difficult problem in itself) and generating a periodic pulse train as excitation also resulted in an unnatural sound.

However, we have observed that the excitation signal, obtained in a CELP coder from the band-limited speech sample, contains low frequency components not present in the input signal. Thus, this excitation signal is used to provide excitation at the low frequencies.

For the coding method mentioned above, we used a stochastic codebook with 512 vectors, an adaptive codebook (see [5]) which allowed for pitch lags from 40 to 140 samples (no fractions), and a gain term for each codebook. All optimum parameters are calculated twice per frame. Since this operation is to be done at the receiver end, the parameters obtained are not quantized.

The synthetic signal generated in this case is only low-pass filtered at 300 Hz and added to $s(n_1)$. The result is an 8kHz sampled signal, band limited to 3.3kHz that contains the recovered low frequencies.

5. RESULTS

We are currently experimenting with single speaker databases. Results are evaluated on sentences outside the training set but within the same speaker.

The results achieved so far for the high frequency reconstruction are illustrated in figure 2, which shows spectrograms for original and reconstructed signals. In the high frequencies, we clearly see the re-occurrence of high energies in fricatives. For unvoiced frames, an over or under-estimation of the envelope slope did not result in perceivable artifacts. However for voiced frames the over-estimation of the spectral envelope slope at high frequencies might be harmful. In figure 3, we show typical examples comparing the predicted and the original envelopes for the high frequency band for unvoiced and

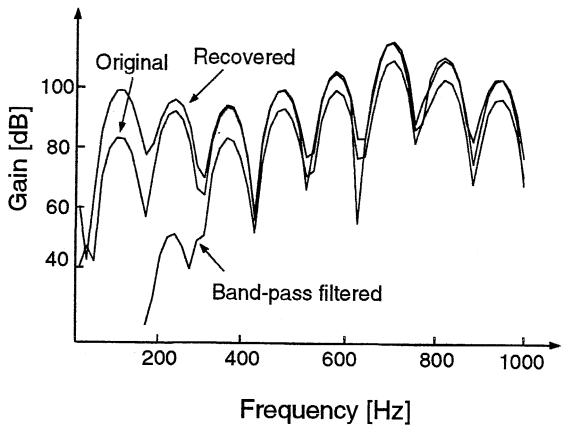

Figure 4: *Power spectrum of original, band-pass filtered, and reconstructed speech (Only the 0-1kHz frequency band is shown). Typical result for a voiced frame*

voiced frames. We can see that the higher formant positions and bandwidths are not always accurately predicted, though the slope and overall shape of the spectral envelopes are consistently well approximated.

For the low frequency recovery, the re-introduction of the fundamental frequency and first harmonics is observed in voiced frames. In figure 4, we show the power spectrum of one voiced frame of the original, band-pass filtered and reconstructed signals. Only the frequency band from 0 to 1 kHz is shown to illustrate the reconstruction of low frequencies.

The technique brings a clear difference in speech quality for both cases. Apart from minor artifacts, the reconstructed speech is quite close to the original high quality speech. The derived mappings appear to be speaker dependent, i.e. the reasonable recovery is possible only within a given speaker. However, further work is necessary to draw definitive conclusions.

6. DISCUSSION

Naturally, we were curious about what were the main reasons for the improved quality of the enhanced speech. In our experiments, we typically have available a) the original envelope, b) the residual of the high quality speech, c) the predicted envelopes and d) the estimated residuals derived from the narrow-band and band-pass filtered samples. This allows us to generate a set of test speech signals which are produced by using different combinations of original values and estimated values.

Our general conclusion from these informal listening tests is: For the low frequency reconstruction, the recovery of the fundamental and lower harmonics of the fine spectral structure was more important than having an accurate envelope shape. For the recovery of high frequencies, the excitation used made almost no difference, as long as the spectral envelope shape was similar to the original.

The two steps taken towards the enhancement of telephone speech have been successful: a broad-band 0-8kHz speech signal has been obtained from a 0-4kHz narrow-band speech signal, and the low frequency band from a 300-3300Hz band-pass filtered speech signal has been approximately reconstructed. However, in order two build a system that will simultaneously recover high and low frequencies from a telephone speech signal, we still need to find a reasonable mapping between the band-limited (300-3300kHz) speech to the higher frequencies. So far, the mapping obtained for this case has been inaccurate and over estimation of the spectral envelope slopes has caused severe artifacts.

7. CONCLUSION

The enhancement of telephone speech by recovery of missing frequencies appears to be viable, and promising results have been obtained. The proposed new technique is relatively simple to implement and future improvements are still under investigation. Although no formal tests have been performed, informal listenings indicate that the subjective quality of narrow-band speech is enhanced by the system.

8. REFERENCES

[1] H. Hermansky, E. Wan, and C. Avendano: Speech enhancement based on temporal processing, *IEEE Proc. ICASSP-95,* pp. 405-408, Detroit 1995.

[2] Y. M. Chen, D. O'Shaughnessy, and P. Mermelstein: Statistical recovery of wideband speech from narrowband speech, *Proc. ICSLP-92*, pp. 1577-1580, Edmonton, Canada 1992.

[3] J.D. Markel and A.H. Gray: *Linear Prediction of Speech,* Springer-Verlag, Germany 1976

[4] V.R. Viswanathan, A.L. Higgings, and W.H. Russel: *Design of a robust baseband LPC coder for speech transmission over 9.6 Kbp/s noisy channel,* IEEE Trans. Comm, VOL.COM.30, NO.4, April 1982

[5] J.P. Campbell, T.E. Tremain, and V.C. Welch: *The DOD 4.8 KBPS standard (proposed federal standard 1016), Advances in Speech Coding,* Kluwer Academic Publishers, 1991.