IMPROVING UPON TOLL QUALITY SPEECH FOR VOIP

Richard V. Cox, David Malah*, David Kapilow

AT&T Labs-Research, *Technion – Israel Institute of Technology

ABSTRACT

It is well known that wider bandwidth speech is preferred both for quality and intelligibility. In this paper we describe a new, low complexity method for creating wider bandwidth speech from clean telephone bandwidth speech. In addition, we describe a packet loss concealment technique that has been standardized for ITU-T Rec. G.711 64 kb/s PCM and is applicable to wider bandwidth speech as well. Together, these two techniques address two of the primary issues with Voice over IP – how to provide greater fidelity to customers and how to overcome packet losses when they do occur.

1. INTRODUCTION

This paper summarizes the results of two projects carried out at AT&T Labs. Their purpose was first to match and then exceed the performance of 64 kb/s G.711 speech that is used in the legacy telephone network throughout the world. In the first portion of the paper we describe our technique for creating wider bandwidth speech from conventional telephone bandwidth speech. In the second portion we address the issue of packet loss concealment (PLC) for both conventional telephony as well as wideband speech. This technique is an ANSI-T1 and ITU standard.

2. CREATING WIDEBAND SPEECH

2.1 *Introduction*

Our motivation is to produce wideband speech from telephone bandwidth speech because it is known that wideband sounds more natural and is generally preferred. Recent work has shown that this is feasible [1-29]. What we wish to achieve is a method that does not rely on training so that the method is robust to channel conditions. At the same time we also want a low complexity system.

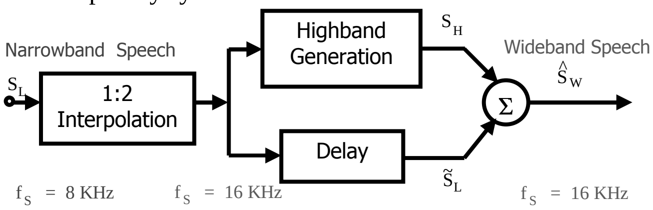

Figure 1 – Typical bandwidth extension system

A typical system is shown in Figure 1. Since the lowband already exists, the input signal needs to be delayed before it can be combined with a synthesized highband signal. In turn the highband signal is typically created by generating a wideband spectral envelope and a wideband excitation and then using only the highband portion of this synthetic signal.

2.2 *The Wideband Envelope*

Speech is produced by a physical system. The well-known acoustic tube model (DATM) of speech production is an analog. Linear prediction analysis produces the parameter values for the model. Our proposal is to interpolate the DATM in order to produce the spectral envelope of the wideband signal. The areas of the DATM correspond to average areas for that region. In order to produce a model that captures twice the bandwidth, we need to interpolate to a model of twice the order. Our approach is to obtain a refinement of the DATM via interpolation followed by re-sampling at the points corresponding to the new section centers. Because the re-sampling is at points that are shifted by a ¼ of the original sampling interval of the underlying vocal tract, we call this process *shifted-interpolation.* Note that with twice as many sections, all the previous section centers are now at section boundaries. We have considered interpolation of the area-ratios and the log-area-ratios. We have also considered zero-order hold, linear interpolation, and cubic spline interpolation in both of these domains. To date, the best results have been obtained using spline interpolation in the log-area-ratio domain.

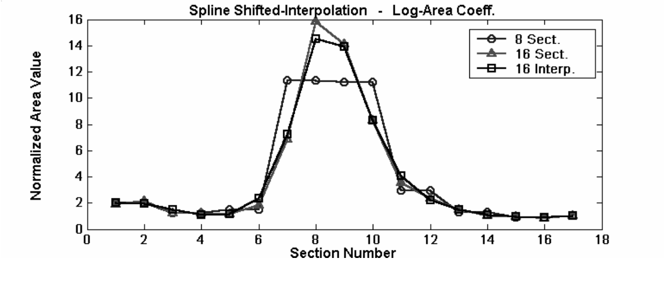

Figure 2 – Area functions for one frame of speech

Figure 2 shows a comparison of the area functions for a 16th order analysis corresponding to one frame of speech. The curve with the triangle points is the actual analysis. If only an 8 kHz signal is analyzed with an 8th order analysis, the curve with the circles results. The spline interpolation in the log area domain produced the curve with the square points.

Figure 3 shows the resulting spectrum comparison for a frame of speech. The thicker line corresponds to the actual spectrum while the thinner line corresponds to the one obtained through spline interpolation in the log-area-ratio domain.

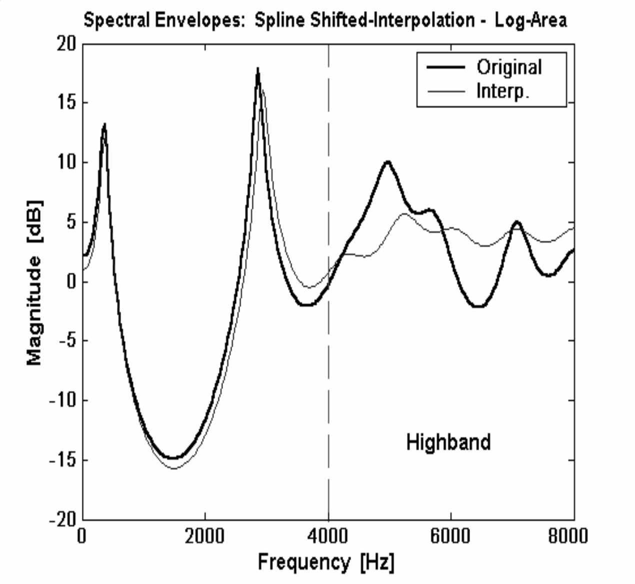

Figure 3 – Comparison of actual and interpolated spectral envelopes.

2.3 *The Wideband Excitation*

We studied halfwave and fullwave rectification of the narrowband excitation signal as possible wideband excitations. To better understand and utilize this method, a mathematical analysis of the spectral characteristics of a signal that models a narrowband residual signal and passes through a generalized wave rectification system was performed. Based on this analysis full-wave rectification was selected and the proper gain needed was computed. A characteristic spectral tilt in the high frequencies was found beneficial and is left uncompensated. Figure 4 is a diagram of the resulting overall system.

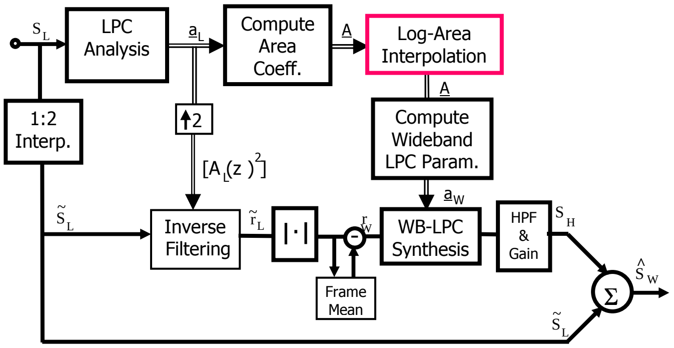

Figure 4 – wideband speech synthesis system

Figure 5 shows two simulation results for the system. Figure 5a compares the composite spectrum for a voiced frame of the original wideband speech and the speech synthesized from narrowband. The original wideband spectra has a dashed line while the synthetic spectra has a solid line. The lower band is identical. Although the upper band is different, it provides the same effect as the actual upper band. The same is true for the unvoiced spectra shown in Figure 5b. Without the original being available for comparison purposes, the synthetic wideband speech sounds natural.

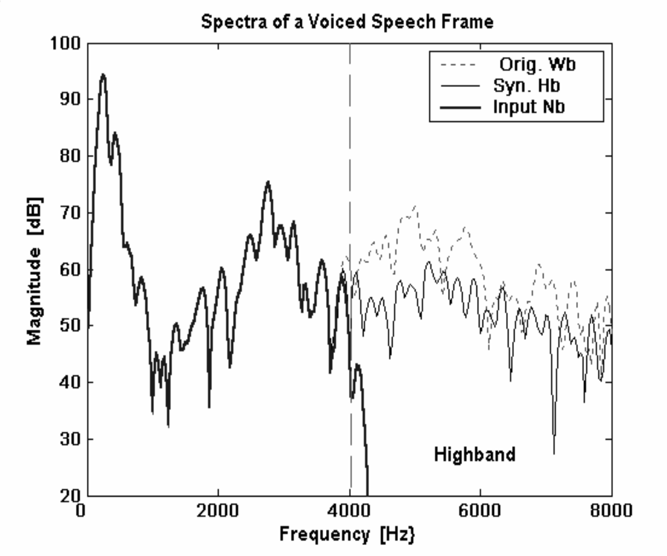

Figure 5a – comparison of spectra for a voiced frame

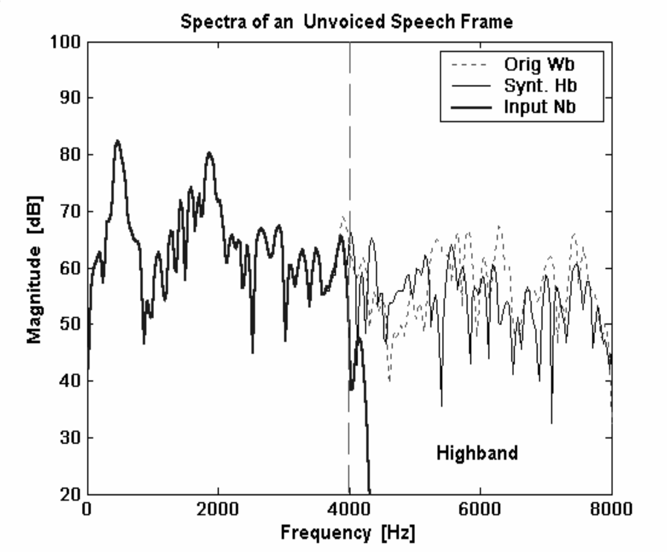

Figure 5b – comparison of spectra for an unvoiced frame

2.4 *Results*

A small subjective test was conducted to determine whether listeners preferred synthetic wideband to narrowband speech. There were 24 sentences. Each comparison was presented in both NB-WB and WB-NB order. The arrangement was random. Subjects were asked to listen to the wideband original followed by the comparison and make an assessment of whether one signal was clearly better, slightly better, or equal in quality to the other. The wideband signal was judged better in 34% of the comparisons, slightly better in 40%, and equal in 10%. We hypothesize that the wideband signal was not preferred in the other comparisons because of small amounts of artifacts that were generated by the wideband excitation signal in some instances. Some further refinement would probably be beneficial.

3. PACKET LOSS CONCEALMENT (PLC)

3.1 *PLC Introduction*

A PLC algorithm hides transmission losses in an audio system where the input signal is encoded and packetized at a transmitter, sent over a network, and received at a receiver that decodes the packet and plays out the output. It is assumed that the receiver has a way of determining if an expected packet does not arrive, arrives too late, or is corrupted in transmission. If a packet is lost or cannot be used, the PLC algorithm hides the missing packet by generating a synthetic packet's worth of audio instead.

Many of the standard ITU-T CELP based speech coders, such as the G.723.1, G.728, and G.729, model speech reproduction in their decoders. These decoders have enough state information to integrate PLC algorithms directly in the decoder, and are specified as part of their standards. G.711, by comparison, is a sample-by-sample encoding scheme that does not model speech reproduction. There is no state information in the coder to aid in the PLC. The PLC algorithm with G.711 is independent of the coder.

The G.711 PLC is a low-complexity time-domain algorithm that uses the most recent 48.75 msec history of the decoded output signal to estimate what the signal should be in the missing frames. The algorithm delays the output by 3.75 msec to allow the synthetic audio signal to be mixed with the tail of the last good packet using an Overlap-Add (OLA) at the start of a loss, insuring a smooth transition between the received and synthetic speech at the loss boundaries. The algorithm has a peak complexity of approximately ½ a MIP that is dominated by a pitch estimate at the start of a loss.

A block diagram of the system with G.711 is shown in Figure 6.

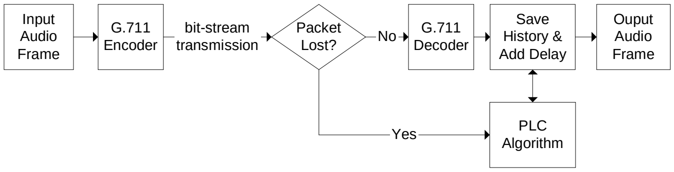

Figure 6 – G.711 Audio Transmission System with PLC

Since the PLC algorithm only depends on the decoded output of G.711, the algorithm will work just as well when no speech coder is present and with wideband speech.

3.2 *PLC Algorithm Overview*

The PLC algorithm conceals the missing packet by generating synthetic speech that has similar characteristics to the speech in the history buffer. The idea is as follows. If the signal is voiced we assume the signal is quasi-periodic and locally stationary. We estimate the pitch using an autocorrelation technique and repeat the last pitch period in the history buffer a few times. A ¼ wavelength OLA is used to smooth each pitch period repetition boundary. However if the loss is long or the pitch period is short (the frequency is high), repeating the same pitch period too many times leads to output that is too harmonic compared with natural speech.

To avoid these artifacts the number of pitch periods used from the history buffer is increased as the length of the loss progresses. Short losses only use the last or last few pitch periods from the history buffer to generate the synthetic signal. Long losses use pitch periods from further back in the history buffer. If the loss exceeds 10 msec, the number of pitch periods used from the history buffer is increased by 1. Beyond 20 msec, another pitch period is added. With long losses the pitch periods from the history buffer are not replayed in the same order that they occurred in the original speech. However, in testing we found the synthetic speech signal generated in long losses still sounds natural.

The longer the erasure, the more likely it is that the synthetic signal will diverge from the lost speech. To avoid artifacts caused by holding certain types of sounds too long the synthetic signal is attenuated as the loss becomes longer. For losses of duration 10 msec or less no attenuation is needed. For losses longer than 10 msec the synthetic signal is attenuated at the rate of 20% per additional 10 msec. Beyond 60 msec, there is not much point in generating the synthetic signal. The estimated signal is so far off that on average it does more harm than good to continue trying to conceal the missing speech after 60 msec, so the output is set to zero (silence).

Whenever a transition is made between signals from different sources it is important that the transition not introduce discontinuities, audible as clicks, or unnatural artifacts into the output signal. These transitions can occur in several places: at the start of the erasure at the boundary between the synthetic signal and the tail of the last good packet, at the end of the loss at the boundary between the synthetic signal and the start of the signal in the first good packet after the loss, when the number of pitch periods used from the history buffer is changed to increase the signal variation, and at the boundaries between the repeated portions of the history buffer. To insure smooth transitions, OLAs are performed at all signal boundaries. Time-domain OLA techniques are known to produce high quality results when time-scaling speech[30,31]. At the boundaries, care is also taken to avoid phase mismatches at the signal's estimated fundamental frequency. The OLAs are weighted with triangular windows to keep the complexity of calculating the variable length windows low. Hanning windows yield similar results.

The previous discussion explains how the algorithm works with stationary voiced speech. If the speech is rapidly changing or unvoiced, the speech may not have a periodic structure. However, by making a few compromises we apply the same algorithm and obtain good results.

First, the smallest pitch period we allow in the pitch estimate is 5 msec, corresponding to a frequency of 200 Hz. While it is known that some high-frequency female and child speakers have fundamental frequencies above 200 Hz, we limit it to 200 Hz so the windows stay relatively large. This way, within a 10 msec lost frame the selected pitch period is repeated a maximum of twice. With high-frequency speakers, this doesn't really degrade the output, since the pitch estimator returns a multiple of the real pitch period. And by not repeating any speech too often, the algorithm does not create periodic speech out of non-harmonic speech. It should be noted that the WSOLA[30] algorithm for time-scaling of speech also uses large OLA windows so the same algorithm can be used with both periodic and non-periodic speech signals.

Second, because the number of pitch periods used to generate the synthetic speech is increased as the erasure gets longer, enough variation is added to the signal that periodicity is not introduced for long erasures.

3.3 *PLC Example*

Figure 7 shows a graphical example of how the PLC algorithm operates. The algorithm's default packet size is 10 msec. The top waveform, labeled "Input" shows the input to the system when 2 consecutive 10 msec packets are lost (a 20 msec loss) in a region of voiced speech from a male speaker.

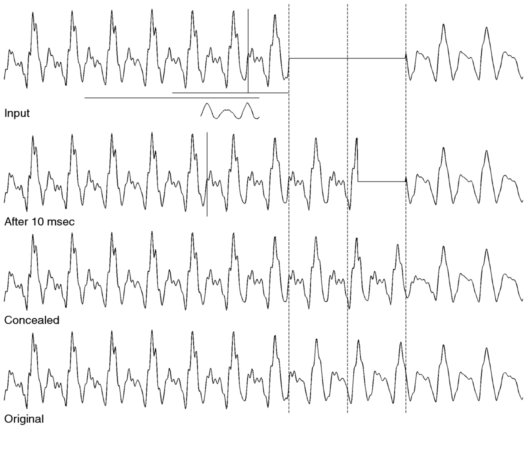

Figure 7 – Packet Loss Concealment Algorithm

Below the Input waveform the normalized autocorrelation used to estimate the pitch at the start of the loss is shown. The top horizontal line corresponds to the 20 msec of speech before the loss. The lower line represents the correlation taps and is a 20 msec window that slides back in time from 5 msec to 15 msec. The output of the correlation is below the horizontal lines. The peak of this graph corresponds to the pitch estimate, shown by the vertical line above the peak.

The 2nd waveform, labeled "After 10 msec" shows the synthetic signal generation after the first 10 msec of the loss. This signal was generated by repeating the last pitch period from the history buffer for the duration of the 10 msec frame. A ¼ wavelength OLA occurs at the start of loss, and when each pitch period is repeated.

The 3rd waveform, labeled "Concealed", shows the complete output of the PLC algorithm after the loss. For the second 10 msec frame, the number of pitch periods used to generate the synthetic signal was increased by 1. An OLA is performed at all transitions, including at the boundary of the received speech in the 1st good packet after the loss.

For comparison purposes the original input signal without a loss is also shown in the waveform labeled "Original". In an ideal system, the concealed speech sounds just like the original. As can be seen in Figure 7, the synthetic waveform closely resembles the original in the missing segments.

3.4 *PLC Results*

The proposed PLC algorithm was compared against techniques for ITU-T Rec. G.728 and G.729 in a formal subjective test. The average results are shown in Figure 8. What these results show is far more robust performance than the other two coders under similar conditions. As a result of its combination of performance, low complexity, and low delay, the proposed algorithm was standardized. It is now Appendix I of ITU-T Rec. G.711 [32] and ANSI Standard T1.521-1999 [33].

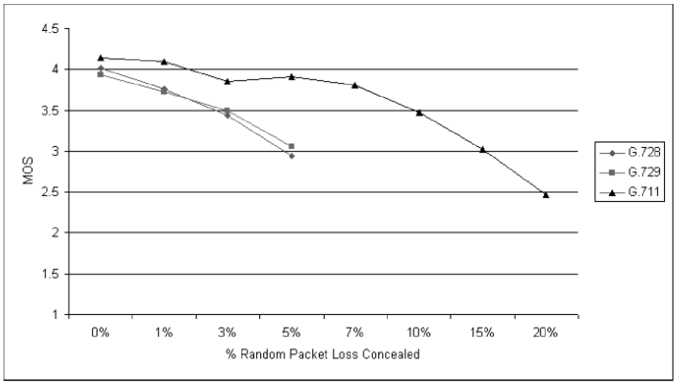

Figure 8 – Subjective Test Results for the G.711 PLC algorithm.

4. REFERENCES

[1] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech," IEEE Trans. Speech and Audio Processing, vol.2, No. 4, pp. 544-548, Oct. 1994.

[2] H. Carl and U. Heute, "Bandwidth Enhancement of Narrow-Band Speech Signals," in Proc. European Signal processing Conf. – EUSIPCO'94, pp. 1178-1181, 1994.

[3] Y. Yoshida and M. Abe, "An Algorithm to Construct Wideband Speech from Narrowband Speech Based on Codebook Mapping," in Proc. Intl. Conf. Spoken Language Processing, ICSLP'94, 1994.

[4] H. Yasukawa, "Quality Enhancement of Band Limited Speech by Filtering and Multirate Techniques," in Proc. Intl. Conf. Spoken Language Processing, ICSLP'94, pp. 1607-1610, 1994.

[5] C. Avendano, H. Hermansky, and E.A. Wan, "Beyond Nyquist: Towards the Recovery of Broad-Bandwidth Speech From narrow-Bandwidth Speech," in Proc. European Conf. Speech Comm. and Technology, EUROSPEECH'95, pp. 165-168, 1995.

[6] H. Hermansky, E.A. Wan, and C. Avendano, "Speech Enhancement Based on Temporal processing," in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'95, pp. 405-408, 1995.

[7] H.Yasukawa, "Enhancement of Telephone Speech Quality by Simple Spectrum Extrapolation Method," in Proc. European Conf. Speech Comm. and Technology, EUROSPEECH'95, 1995.

[8] C-F. Chan and W-K. Hui, "Wideband Re-Synthesis of narrowband Celp-Coded Speech using Multiband Excitation Model, in Proc. Intl. Conf. Spoken Language Processing, ICSLP'96, pp. 322-325, 1996.

[9] H. Yasukawa, "Restoration of Wide Band Signal from Telephone Speech Using Linear Prediction Error Processing," in Proc. Intl. Conf. Spoken Language Processing, ICSLP'96, pp. 901-904, 1996.

[10] H. Yasukawa, "Adaptive Filtering for Broad Band Signal Reconstruction using Spectrum Extrapolation," in Proc. IEEE Digital Signal Processing Workshop, pp. 169-172, 1996.

[11] H. Yasukawa, "A Simple Method of Broad Band Speech Recovery from Narrow Band Speech for Quality Enhancement," in Proc. IEEE Digital Signal Processing Workshop, pp. 173-175, 1996.

[12] H. Yasukawa, "Restoration of Wide Band Signal from Telephone Speech using Linear Prediction Residual Error Filtering," in Proc. IEEE Digital Signal Processing Workshop, pp. 176-178, 1996.

[13] H. Yasukawa, "Implementation of Frequency Domain Digital Filter for Speech Enhancement," in Proc. Intl. Conf. Electronics, Circuits and Systems, ICECS'96, pp. 518-521, 1996.

[14] H. Yasukawa, "Signal Restoration of Broad Band Speech Using Nonlinear Processing," in Proc. European Conf. Speech Comm. and Technology, EUROSPEECH'96, pp. 987-990, 1996.

[15] Y. Nakatoh, M. Tsushima, and T. Norimatsu, "Generation of Broadband Speech from Narrowband Speech using Piecewise Linear Mapping," in Proc. European Conf. Speech Comm. and Technology, EUROSPEECH'97, 1997.

[16] J. Epps and W.H. Holmes, "Speech Enhancement using STC-Based Bandwidth Extension," in Proc. Intl. Conf. Spoken Language Processing, ICSLP'98, 1998.

[17] H. Yasukawa, "Wideband Speech recovery from Bandlimited Speech in Telephone Communications," in Proc. Intl. Symp. Circuits and Systems, ISCAS'98, pp IV-202 – IV-205, 1998.

[18] J. Epps and W.H. Holmes, "A New Technique for Wideband Enhancement of Coded Narrowband Speech," in Proc. IEEE Speech Coding Workshop, SCW'99, 1999.

[19] N. Enbom and W.B. Kleijn, "Bandwidth Expansion of Speech based on Vector Quantization of the Mel Frequency Cepstral Coefficients," in Proc. IEEE Speech Coding Workshop, SCW'99, 1999.

[20] P. Jax and P. Vary, "Wideband Extension of Telephone Speech using Hidden Markov Model," in Proc. IEEE Speech Coding Workshop, SCW'00, 2000.

[21] J-M. Valin and R. Lefebvre, "Bandwidth Extension of Narrowband Speech for Low Bit-Rate Wideband Coding," in Proc. IEEE Speech Coding Workshop, SCW'00, 2000.

[22] K-Y. Park and H.S. Kim, "Narrowband to Wideband Conversion of Speech using GMM Based Transformation," in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'00, pp. 1843-1846, 2000.

[23] G. Miet, A. Gerrits, and J.C. Valiere, "Low-Band Extension of Telephone-Band Speech," in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'00, pp. 1851-1854, 2000.

[24] S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, "Speech Enhancement via Frequency Bandwidth Extension using Line Spectral Frequencies," in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'01, 2001.

[25] A. Uncini, F. Gobbi, and F. Piazza, "Frequency Recovery of Narrow-band Speech using Adaptive Spline Neural Networks," in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'99, 1999.

[26] P. Jax and P. Vary, "On Artificial Bandwidth Extension of Telephone Speech," Signal Processing, Vol. 83, pp. 1707-1719, 2003.

[27] G. Chen and V. Parsa, "HMM-based Frequency Bandwidth Extension for Speech Enhancement using Line Spectral Frequencies", in Proc. Intl. Conf. Acoust., Speech, Signal processing, ICASSP'04, pp. I-709 - I-712, 2004.

[28] S. Jaisimha and I. Y. Soon, "Bandwidth Extension of Narrow Band Speech using Cepstral Linear Prediction", in Proc. Intl. Conf. Information, Communications and Signal Processing, ICICS'03, pp. 1404-1407, 2003.

[29] H. Gustafsson and I. Claesson, "Speech Bandwidth Extension", in Proc. Int. Conf. Multimedia and Expo, ICME'01, pp.1016-1019, 2001.

[30] W. Verhelst and M. Roelands, "An Overlap-Add Technique based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech", in Proc. Intl. Conf. Acoust., Speech, Signal Processing, ICASSP'93, vol 2, pp. 554-557, 1993.

[31] E. Moulines and F. Charpentier, "Pitch Synchornous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones", Speech Comm., vol 9, no. 5/6, pp. 453-467, 1990.

[32] "A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711," Recommendation G.711 Appendix I (09/99), International Telecommunications Union.

[33] "Packet Loss Concealment for Use with ITU-T Recommendation G.711," ANSI Standard T1.521-1999, American National Standards Institute, Dec. 1999.