LOSS RECOVERY THROUGH SPECTRAL INTERPOLATION FOR ROBUST SPEECH RECOGNITION OVER PACKET VOICE COMMUNICATIONS

Amr H. Nour-Eldin, Hesham Tolba and Douglas O'Shaughnessy

INRS-ÉMT, Université du Québec, Montréal, Québec, Canada
{nour,tolba,dougo}@inrs-emt.uquebec.ca

ABSTRACT

Packet voice communications generally suffer packet losses as a result of various network- or transmission-related impairments. Upon decoding, these lost packets result in missing speech segments that degrade *automatic speech recognition* (ASR) performance. We present a novel *loss recovery* scheme that reproduces the missing speech waveform by interpolating its spectrum from the speech spectra on both sides of a loss. An adaptive mechanism is used to determine the FFT width of the speech waveform before and after a loss to capture as much spectral detail as possible. A linearly weighted *spectral interpolation* ensues to obtain the spectra of missing speech. The missing speech waveform is then reconstructed through IFFT, followed by smoothing at packet boundaries. Tests on Bluetooth voice packets with a high loss rate of 38% show that our scheme improves ASR performance considerably (up to 20%) while being computationally efficient, as it is an FFT-based scheme.

1. INTRODUCTION

A well-known problem in packet voice communications is that of packet losses which occasionally occur as a result of various impairments. Network-related impairments include misrouting, excessive delay, and congestion. The affected voice packets are either lost or arrive at the destination too late to be of any use. Furthermore, adverse transmission conditions may result in too many uncorrectable bit errors in the transmitted voice packets rendering these packets unacceptable to the receiver, and are, hence, discarded. Such is the case for wireless packet voice communications (e.g., GSM, Bluetooth), which are susceptible to many radio frequency (RF) channel impairments (e.g., interference, fading, AWGN).

Upon decoding, these packet losses generally manifest themselves as missing speech segments that may severely degrade speech quality and ASR performance depending on the packet loss rate and packet lengths. For codecs exploiting speech redundancies to decrease the overall bit-rate where the correct decoding of speech packets depends on preceding ones (e.g., adaptive/differential PCM, adaptive delta modulation), a lost packet additionally affects subsequent ones adversely. An example is the erroneous scaling of decoded Bluetooth CVSD voice packets following a packet loss [1]. Hence, a packet *loss recovery* or *error concealment* scheme is typically used to recover the missing speech segments or to fill in the gaps in the output speech signal. Although much research has been directed towards the problem of packet loss recovery, this research has generally been concerned with the improvement of subjective speech quality [2]–[7] rather than ASR performance. While some have examined the effects of packet losses on speech intelligibility [8, 9] (typically defined as the number of words correctly heard [10]), the correlation between subjective intelligibility and machine ASR has not been clearly defined. Furthermore, very few [1, 11, 12, 13] have directly investigated the effects of packet losses on ASR performance or attempted to alleviate such effects. Accordingly, we present an interpolation-based loss recovery scheme with the objective of ASR performance improvement. However, as it is a flexible loss recovery scheme which also performs speech waveform reconstruction (contrary to those of [11]–[13]), it can be generalized and applied to speech quality improvement as well. Hence, a review of interpolation-based schemes is presented in Section 2.

Packet loss recovery techniques are generally classified into sender-based and receiver-based schemes. A survey of such techniques can be found in [2]. In contrast to sender-based schemes, receiver-based ones require no modifications to packet format or to a transmission system. They rely on producing a replacement for a lost packet which is as similar to the original as possible, with no assistance from the sender. This is possible since speech exhibits large amounts of short-term self-similarity. Their ability to add any improvement depends greatly on several factors (e.g., loss rate, packet and loss lengths, phonetic importance of loss locations, etc.) [1, 2].

In order of increasing performance and complexity, receiver-based schemes are classified into *insertion-*, *interpolation-*, and *regenerative-based* schemes, consecutively. The computation required to perform the more advanced repair techniques increases greatly relative to the simpler repair options, while the improvement in quality achieved by such schemes is incremental at best [2]. Thus, interpolation-based schemes represent a good compromise between the simpler but less-performing insertion-based schemes and the higher-performing but demanding regenerative-based ones.

The work presented in this paper is part of our recent research on the robust recognition of Bluetooth speech in the presence of RF interference. We concluded in [1] that packet losses are primarily responsible for the degradation in ASR performance. Accordingly, we proposed several modifications to Bluetooth's CVSD decoder such that insertion-based loss recovery is implemented. Measures were also taken in order to correct step-size errors in packets following a loss. Although our results showed that ASR performance can be considerably improved through insertion-based schemes, we sought further improvements through efficient interpolation-based loss recovery (since Bluetooth is a low-power communication standard). Thus, we begin by reviewing interpolation-based loss recovery in Section 2. In Section 3, we present our scheme. Section 4 details our ASR results and conclusions are given in Section 5.

2. INTERPOLATION-BASED PACKET LOSS RECOVERY

Interpolation-based techniques attempt to interpolate from packets surrounding a loss to produce a replacement for the lost packet, thus accounting for the changing characteristics of a signal. In the well-known *waveform substitution* [3], a replacement for a lost speech packet is searched for in the speech preceding (one-sided) or on both sides of the loss (two-sided) using pattern matching to a template immediately preceding the loss. The replacement packet is repeated for multiple-packet losses. Two-sided schemes generally perform better than one-sided ones in terms of subjective speech quality [4]. Another waveform substitution technique based on pitch detection, proposed in [3], additionally performs pitch detection on previous speech samples. Losses during unvoiced speech segments are repaired using packet repetition, and voiced losses repeat a waveform of appropriate pitch. This technique, known as *pitch waveform replication*, was found in [4] to perform marginally better than the pattern matching-based approach.

Two assumptions are implicit in the waveform substitution approach, making it flawed. First, it assumes that there will always be enough samples between the best template match and the loss-onset from which to extract the needed samples to replace a lost packet. This assumption is invalid since a best match may be found

as close to the template as 2 ms for a highly-pitched ($F_0 = 500$ Hz) voiced female vowel for example, whereas packet lengths are typically longer, leaving the rest of the lost packet(s) unaccounted for. Secondly, even in the case where a best match is found such that there are enough samples to replace a lost packet, the repetition of these samples for multiple-packet losses is not justified since those samples do not correspond to the periodicity of speech at this time interval if such speech is in fact periodic (voiced). On the other hand, as stated in [3], if the pre-loss speech is non-periodic (unvoiced), the repetition of samples will "create a highly periodic signal" in place of the non-periodic missing speech, and hence, distorting the unvoiced characteristic of speech in this time interval. This led the authors to conclude that their proposed waveform substitution techniques can only improve speech quality up to packet loss rates of 0.3, with the quality breaking down for higher rates.

For a given perceptual quality, an improved but more computationally expensive technique [5] increases the loss lengths to about twice of what can be tolerated when using pitch detection alone, by using a phase-matching reconstruction scheme to ensure phase continuity between the replacement and the ensuing speech.

A recent promising technique operating on PCM speech [6], extracts the residual signal of the previously received speech by linear prediction analysis, uses periodic replication to generate an approximation for the excitation signal of the missing speech, and generates synthesized speech using this excitation. Although this algorithm was found to be better than those employing pitch detection, the additional complexity introduced (≈2–5 times higher) makes it less suitable for implementation in resource-limited environments.

Time scale modification allows speech on either side of a loss to be stretched across the loss. Sanneck *et al.* [7] present a scheme that finds overlapping vectors of pitch cycles on either side of the loss, offsets them to cover the loss and averages them where they overlap. Although the technique appears to work better than waveform substitution and pitch waveform replication, it is more demanding.

As noted previously, these schemes—like most loss recovery schemes—are concerned with the improvement of subjective speech quality and were not tested or applied for ASR purposes. In terms of recognition-oriented interpolation-based loss recovery, Milner [11, 12] recently proposed a scheme that operates on coded packets containing log-filterbank or MFCC features rather than PCM speech, performing interpolation in the log-filterbank domain. Although this technique can recover lost packets with good accuracy up to 50% packet loss where it nearly halves the ASR error rate, it has several drawbacks. As it operates on specifically coded voice packets, this technique can not be applied to most networks, which typically employ PCM or LP-derived coding for voice packets. Moreover, since speech reconstruction from filterbank-based features is not possible, this technique is limited to recognition purposes. It further excludes the application of most speech enhancement algorithms—which operate on PCM speech—prior to recognition to partly remove additive acoustic noise, which aids in the recognition process. Moreover, it is based on the assumption that each packet contains one feature vector. This is very inefficient since most ASR systems use speech windows of 10–25 ms to obtain feature vectors. Networks whose packet lengths are outside this range can not use this technique without requiring a change of the ASR system. Even if ASR adapting is performed, the resulting performance would depend greatly on packet length. For short packets ($\leq$ 10 ms), the obtained features would not have sufficient frequency resolution to provide phonetically discriminating vectors, and consequently ASR performance would likely not improve. If the packets were to be longer, the degradation effect of lost packets on subsequent ASR performance would increase since longer packets inherently mean longer durations of lost speech, making the interpolation process more susceptible to error. Finally, the results reported in [11, 12] are based on a limited digit recognition task—a small-size vocabulary which does not consider practical medium- or large-size vocabulary tasks. These arguments also apply for the scheme presented in [13]. In contrast, the spectral interpolation scheme we present next addresses these problems.

3. SPECTRAL INTERPOLATION

We propose a spectral interpolation scheme that reproduces the missing speech waveform during a loss duration by interpolating its spectrum from the speech spectra on both sides of the loss. To capture as much spectral detail as possible to use for the interpolation process, an adaptive mechanism is used to determine the FFT width of the speech waveform on either side of a loss based on the time interval between the loss and the following one with a controllable upper limit. A linearly weighted spectral interpolation ensues to obtain the spectra of missing speech, with increased weighting for the pre- or post-loss spectrum depending on the relative nearness of the lost speech duration whose spectrum is being reconstructed. This weighting technique incorporates the dynamic speech characteristics before and after a loss into the reconstruction process. The missing speech waveform is then reconstructed using inverse FFT, followed by smoothing at boundaries to remove discontinuities.

This scheme is similar to those of [6, 11, 12] in the aspect of exploiting the slowly varying spectral characteristics of speech. However, while that of [6] is one-sided, our scheme is two-sided, and hence, can better account for variations in speech characteristics. Although a delay is also introduced, our scheme is flexible in the sense that it allows control over the amount of this delay at the expense of spectral detail. Moreover, rather than using linear prediction and pitch estimation as in [6], using FFT-based spectral analysis is much more efficient on most current signal processors. This also makes it more efficient than waveform substitution techniques [3]. The use of spectral coefficients, as opposed to filterbank-based ones as in [11, 12], further allows speech reconstruction, thus allowing the application of speech enhancement algorithms for robust ASR or subjective quality improvement. Since our scheme operates on PCM speech, it is essentially network-independent since voice packets (whether PCM or coded) are eventually decoded at the destination. Contrary to [11]–[13], our scheme makes no assumptions about packet lengths or their contents. Depending on the available information around a loss, it rather treats a loss duration as part of a bigger frame (useful for short packet losses), as a single frame, or as a contiguous length which can be divided into several frames (useful for long packet losses). Whichever course the scheme takes, it is determined regardless of packet length or the exact number of packets within a loss, but rather in a way such that the best possible use is made of the available information surrounding the loss. This great flexibility allows the application of our scheme in almost any type of packet voice network. The only assumption we make here is that a loss detection scheme is readily available. This assumption is practically realized in most networks, which typically employ an acknowledgment or error detection scheme.

The spectral interpolation scheme proceeds in real-time by using the output of the receiver's loss detection scheme as a control signal. Upon reception of a packet loss signal at time index n_0, the speech decoder's output waveform samples $y(n)$ of the previous P_{max} packets (where 1 packet contains N waveform samples), i.e.,

$$y_{pre}(n) \triangleq y(n), \qquad n_0 - (P_{max}N) \leq n < n_0, \tag{1}$$

are stored in a *pre-loss buffer*. Part of the stored waveform samples may have been reconstructed by a previous interpolation step.

Using the input control signal, the loss length represented by the number of consecutively lost packets Q, is continuously updated until correct reception of a packet occurs. The ensuing decoded output waveform samples are then stored in a *post-loss buffer* until a new loss occurs or the number of stored waveform packets P reaches a maximum of P_{max}, i.e., $1 \leq P \leq P_{max}$. Thus,

$$y_{post}(n) \triangleq y(n), \qquad n_0 + Q \cdot N \leq n < n_0 + (Q+P)N. \tag{2}$$

Hence, for a sampling rate of F_s, the end-to-end delay introduced in this scheme is given by

$$(Q_{max} + P_{max})\frac{N}{F_s}. \tag{3}$$

Based on P, the number of stored post-loss waveform packets, two Fourier transform operations of width $P \cdot N$ are performed on the $P \cdot N$ pre- and post-loss waveform samples immediately before and after the loss, as given by

$$Y_{pre}(k) = \sum_{\hat{n}=0}^{PN-1} y_{pre}(\hat{n} + n_0 - PN) e^{-j(2\pi/(PN))\hat{n}k}, \quad (4a)$$

$$Y_{post}(k) = \sum_{\tilde{n}=0}^{PN-1} y_{post}(\tilde{n} + n_0 + QN) e^{-j(2\pi/(PN))\tilde{n}k}, \quad (4b)$$

where $\hat{n} \triangleq n - (n_0 - PN)$, $\tilde{n} \triangleq n - (n_0 + QN)$, and k is the frequency index. The obtained spectra are then linearly interpolated depending on the number of lost packets Q and the number of waveform packets P used in the transform operations of Eq. (4) using multiplicative spectral weighting factors. The purpose of varying the multiplicative interpolation factors is to weight each component pre- or post-loss spectrum according to the location of the speech interval whose spectrum is being reconstructed through interpolation. In other words, the closer the missing speech interval is to the pre-loss (or post-loss) waveform, the closer its spectrum should resemble that of the pre-loss (or post-loss) spectrum by increasing the weighting factor of the pre-loss (or post-loss) spectrum in the interpolation process. We note here that we use the phase of the complex spectrum obtained by interpolating the pre- and post-loss complex spectra directly as the phase of the missing speech interval. Although a phase-matching reconstruction scheme as that of [5] may provide a slightly better replacement in terms of perceptual quality, such schemes are computationally expensive. In addition to the fact that ASR systems typically ignore phase properties [10] and the unimportance of phase in objective (SNR) speech enhancement [14], we accordingly decided that simple smoothing at replacement boundaries (as described below) is sufficient and that further complicated phase processing is uncalled for.

For a time frame index l, the interpolation process takes one of four possible forms based on the lengths of P and Q relative to each other as follows:

(1) $P = Q$

We assume that speech in the interval before, during, and after the loss, i.e., $n_0 - PN \le n < n_0 + (Q+P)N$, is divided into three *non-overlapping* rectangular frames of length PN into $Y_{pre}(k, l-1)$, $Y(k,l)$, and $Y_{post}(k, l+1)$, respectively. In this case the pre- and post-loss spectra, $Y_{pre}(k, l-1)$ and $Y_{post}(k, l+1)$, respectively, are averaged equally by

$$\hat{Y}(k,l) = \frac{1}{2} Y_{pre}(k, l-1) + \frac{1}{2} Y_{post}(k, l+1), \quad (5)$$

where $\hat{Y}(k,l)$ represents an estimate of the loss interval spectrum at frequency index k. The reconstructed speech waveform is then obtained by

$$\hat{y}(n) = \sum_{k=0}^{PN-1} \hat{Y}(k,l) e^{j(2\pi/(PN))k(n-lPN)}, \quad n_0 \le n < n_0 + QN. \quad (6)$$

To ensure continuity at the replacement waveform boundaries, an efficient cubic interpolation operation [15] is performed where a cubic polynomial is fitted between the two outer samples at $(n_0 - M/2 - 1)$ and $(n_0 + M/2)$ for the loss-onset boundary, and between $(n_0 + QN - M/2 - 1)$ and $(n_0 + QN + M/2)$ for the loss-end boundary, where M is an even number of boundary samples to interpolate. The interpolated sample values are calculated such that the first derivative $y'(n)$ is continuous and the slopes at the two end points are "shape-preserving" and "respect monotonicity", thus ensuring waveform shape preservation and continuity at replacement waveform boundaries.

(2) $P > Q$

In this case, we assume that speech has been divided into three *overlapping* rectangular frames of length PN. Spectral interpolation is performed according to Eq. (5) and waveform reconstruction is similarly performed by Eq. (6). However, the reconstructed waveform now overlaps the pre- and post-loss waveforms. To effect continuity at the replacement speech boundaries, a simple Hanning weighted overlap-and-add operation is performed on both edges using M samples at each boundary. Thus, only the $QN + 2M$ intermediate samples obtained from Eq. (6) are used for replacing the QN missing waveform samples and the $2M$ boundary samples required for the smoothing operations, i.e.,

$$\hat{y}(n) = \sum_{k=0}^{PN-1} \hat{Y}(k,l) e^{j(2\pi/(PN))k(n-lPN)}, \quad n_0 - M \le n < n_0 + QN + M. \quad (7)$$

(3) $P < Q,\ Q \bmod P = 0$

If Q is an integer number of P such that $Q = bP$ where b is an integer greater than 1, the speech interval from $n_0 - PN$ to $n_0 + (Q+P)N$ spanning the pre-loss, lost packets, and post-loss durations is assumed to be divided into $b+2$ *non-overlapping* rectangular frames each of size PN. Let $a = \frac{1}{1+b}$ be a *weight-unit*, then the pre- and post-loss spectra, which are now represented by $Y_{pre}(k,l)$ and $Y_{post}(k, l+b+1)$, respectively, are interpolated by

$$\hat{Y}(k, l+m) = (1 - (m \cdot a)) Y_{pre}(k,l) + (m \cdot a) Y_{post}(k, l+b+1), \quad (8)$$

where $1 \le m \le b$ is an integer multiplicative coefficient corresponding to the time index order of the missing speech frame whose spectrum is being reconstructed by interpolation. Eq. (8) thus effects a linearly weighted interpolation, in which the weights depend on the position of the missing speech frame being reconstructed among all the missing frames. This incorporates changes in speech characteristics before and after a loss into the interpolation process. Finally, the missing speech waveform estimate is obtained by

$$\hat{y}(n) = \sum_{m=1}^{m=bPN-1} \sum_{k=0} \hat{Y}(k, l+m) h\big(n - (l+m-1)PN\big) \cdot e^{j(2\pi/(PN))k(n-(l+m-1)PN)}, \quad (9)$$

where $h(n)$ is a rectangular window of length PN, such that $h(n) = 1$ for $0 \le n < PN$ and $h(n) = 0$ otherwise. Waveform continuity at loss boundaries is effected through a cubic interpolation operation similar to that described above.

(4) $P < Q,\ Q \bmod P \neq 0$

In this case, the number of pre- and post-loss waveform samples to be used for interpolation, and consequently the Fourier transform widths, is varied from PN such that the number of missing speech samples to be reconstructed by interpolation, QN, is the minimum integer multiple of that number. Thus, if N' is the required number of interpolation samples, then it should satisfy $QN = bN'$, where b is the minimum integer satisfying this relation. We search for the minimum possible integer since we desire to use as many pre- and post-loss samples as possible, thereby incorporating more spectral detail in the interpolation process. The minimum b is searched for by incrementing b in steps of 1, starting with a value of 2, until it satisfies the conditions $(QN/b) < PN$ and $(QN/b) \bmod 1 = 0$, with the latter condition ensuring that QN/b is an integer number of samples. Accordingly, the number of pre- and post-loss waveform samples to be interpolated is $N' = QN/b$.

With PN being substituted by N', Fourier transforms of width N' are performed to give $Y_{pre}(k)$ and $Y_{post}(k)$ according to Eqs. (4), followed by interpolation in a manner similar to that of the previous case (case 3) using Eq. (8), and finally reconstructing the missing speech using Eq. (9), followed by a cubic interpolation boundary smoothing operation.

4. ASR RESULTS

The interpolation accuracy and the delay incurred in our scheme depend on two controllable factors: the number of boundary samples to be smoothed, M; and the maximum number of packets surrounding a loss to use for interpolation, P_{max}. While the optimal value for M can only be determined empirically, it does not affect the overall delay of the scheme. In contrast, P_{max} can be varied for better interpolation or for shorter delay. As shown by Eq. (3), a higher P_{max} involves a longer delay. For short voice packets, P_{max} should be large enough to capture sufficient spectral detail while at the same time not exceeding by much the 10–15 ms quasi-stationarity of speech during which speech properties can be considered stationary [10]. For example, Bluetooth HV3 (High quality Voice) packets are 3.75 ms long, thus requiring that P_{max} be in the range of 2–5 packets. On the other hand, P_{max} should not exceed 2 for long packets ($>$ 15 ms). The other parameter affecting delay in Eq. (3), i.e., Q_{max}, depends on the packet loss rate. However, since $P(Q=q)$ is typically an exponential function of the loss rate r, i.e. $P(Q=q)=r^q$, Q_{max} rarely exceeds 4 or 5 packets even at high loss rates. For example, at a loss rate of 50%, $P(Q=5)=0.5^5=0.031$.

Thus, through careful setting of P_{max}, our scheme can handle most packet sizes. Moreover, as it also deals efficiently with low loss rates (through cases 1 and 2) as well as high loss rates involving multiple-packet losses (cases 3 and 4), this scheme can be applied to virtually any type of network with PCM output voice.

In the context of our research on Bluetooth speech, the spectral interpolation scheme was applied to the output speech waveforms of several Bluetooth CVSD decoders employing insertion-based loss recovery of HV3 packets with a high loss rate of 38.6% [1]. The use of waveforms on both sides of losses in spectral interpolation loss recovery is justified by the step-size correction measures implemented in the Bluetooth CVSD decoder [1]. The system used for automatic recognition is a speaker-independent triphone HMM-based recognizer using MFCC parameters. It was constructed for the *medium-size* continuous speech TIMIT task of 6146 words, giving a word recognition rate of 98.6%. The Bluetooth CVSD decoders and the ASR system are described in more detail in [1]. It should be noted here that a comparison with other recognition-oriented loss recovery schemes (e.g., [11]–[13]) is not feasible since such schemes, as previously described, do not operate on PCM speech but rather on specifically coded voice packets.

As Bluetooth HV3 packets are 3.75 ms long, and since our analysis showed that 97% of packet losses are $\leq$ 15 ms, even at a 38% loss rate, P_{max} was set to 4, i.e., 15 ms. Q_{max} was found empirically to be 7, and hence, the overall delay given by Eq. (3) is 41.25 ms.

Table 1 lists the ASR results for the seven Bluetooth CVSD decoders of [1] with and without spectral interpolation. $M=2,\ 4$, and 6 boundary samples were used for the smoothing operations. The results of Table 1 show that using our scheme will always improve performance (except for model 7) regardless of M and the high loss rate, with the improvement being quite significant in some cases (e.g., model 5). ASR performance improvement ranges from 4.1% for model 3 to a considerable 20.5% for model 5. Table 1 also shows that the number of samples M used for boundary smoothing slightly affects ASR performance improvement. In general, the best results were obtained using $M=2$ or 4 samples. Finally, results show that spectral interpolation loss recovery clearly outperforms insertion-based techniques while being computationally efficient.

Bluetooth Model	Without Interpolation	With Spectral Interpolation		
		$M=2$	$M=4$	$M=6$
1	76.42	83.37	83.94	81.84
2	80.11	83.30	84.26	81.71
3	79.99	83.24	82.73	79.92
4	76.29	82.66	81.13	78.20
5	65.77	77.82	78.90	79.22
6	73.74	78.39	78.71	79.54
7	81.52	77.06	79.09	78.27

Table 1: ASR performance (in % word recognition rate) using the spectral interpolation scheme with different values for M.

5. CONCLUSION

We present a novel loss recovery scheme for robust speech recognition over packet communications. It exploits the slowly varying characteristics of speech through FFT-based spectral interpolation, and hence, is more efficient than other interpolation-based schemes. Tests on Bluetooth speech with a high loss rate show that our scheme can improve ASR performance considerably (up to 20%) over traditional insertion-based schemes. Contrary to other recognition-oriented loss recovery schemes, it is a flexible network-independent scheme that can be adapted to efficiently handle most packet sizes and loss rates, particularly multiple-packet losses. Furthermore, it performs speech reconstruction, and hence, allows the implementation of most speech enhancement algorithms for the purposes of robust ASR as well as subjective quality improvement.

REFERENCES

[1] A. H. Nour-Eldin, H. Tolba and D. O'Shaughnessy, "Automatic recognition of Bluetooth speech in 802.11 interference and the effectiveness of insertion-based compensation techniques", to appear in *Proc. ICASSP*, 2004.

[2] C. Perkins, O. Hodson and V. Hardman, "A survey of packet-loss recovery techniques for streaming audio", *IEEE Network Mag.*, vol. 12, no. 5, pp. 40–48, 1998.

[3] D. J. Goodman, G. B. Lockhart, O. J. Wasem and W. C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications", *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 34, no. 6, pp. 1440–1448, 1986.

[4] O. J. Wasem, D. J. Goodman, C. A. Dvorak and H. G. Page, "The effect of waveform substitution on the quality of PCM packet communications", *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 36, no. 3, pp. 342–348, 1988.

[5] R. A. Velenzuela and C. N. Animalu, "A new voice-packet reconstruction technique", in *Proc. ICASSP*, vol. 2, pp. 1334–1336, 1989.

[6] E. Gündüzhan and K. Momtahan, "A linear prediction based packet loss concealment algorithm for PCM coded speech", *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 778–785, 2001.

[7] H. Sanneck, A. Stenger, K. Ben Younes and B. Girod, "A new technique for audio packet loss concealment", in *Proc. GLOBECOM*, vol. 1, pp. 48–52, 1996.

[8] R. M. Warren, *Auditory Perception*, Pergamon Press, 1982.

[9] R. C. F. Tucker and J. E. Flood, "Optimizing the performance of packet-switched speech", in *Proc. IEEE Conf. Digital Process. of Signals in Commun.*, pp. 227–234, 1985.

[10] D. O'Shaughnessy, *Speech Communications — Human and Machine*, IEEE Press, 2000.

[11] B. Milner and S. Semnani, "Robust speech recognition over IP networks", in *Proc. ICASSP*, vol. 3, pp. 1791–1794, 2000.

[12] B. Milner, "Robust speech recognition in burst-like packet loss", in *Proc. ICASSP*, vol. 1, pp. 261–264, 2001.

[13] C. Peláez-Moreno, A. Gallardo-Antonlin and F. Diaz de Maria, "Recognizing voice over IP: A robust front-end for speech recognition on the World Wide Web", *IEEE Trans. Multimedia*, vol. 3, pp. 209–218, 2001.

[14] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement", *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. 30, no. 4, pp. 679–681, 1984.

[15] F. N. Fritsch, R. E. Carlson, "Monotone piecewise cubic interpolation", *SIAM J. Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980.