INTERPOLATION OF LOST SPEECH SEGMENTS USING LP-HNM MODEL WITH CODEBOOK-MAPPING POST-PROCESSING

Esfandiar Zavarehei Saeed Vaseghi

Department of Electronic and Computer Engineering, Brunel University, London, UK

{esfandiar.zavarehei, saeed.vaseghi}@brunel.ac.uk

ABSTRACT

This paper presents a method for interpolation of lost speech segments. The short-time spectral amplitude (STSA) of speech is modeled using a linear prediction (LP) model of the spectral envelop and a harmonic plus noise model (HNM) of the excitation. The restoration algorithm is based on interpolation of the parameters of LP-HNM models of speech from both side of the gap. A codebook mapping (CBM) technique is used to fit the interpolated parameters to a pre-trained speech model. Experiments show that the CBM module mitigates the artifacts that may result from interpolation of relatively long speech gaps. Evaluations demonstrate that the proposed interpolation method results in a superior quality in comparison to alternative restoration methods.

1. INTRODUCTION

This paper describes a model-based signal interpolation method for restoration of lost speech segments. The interpolation method can be used in a number of applications for estimation of speech segments that are missing or lost to noise or dropouts such as for packet loss concealment (PLC) in speech communication over mobile phones or voice over IP (VoIP), for restoration of archived speech recordings and for general purpose interpolation.

In general, speech interpolation methods utilize signal models that capture the correlations of speech parameters on both sides of the missing speech segment. Algorithms specifically designed for speech gap restoration can be categorized into two classes, i) predictive or extrapolative, where only the past samples are available and ii) estimative or interpolative, where some future samples are also available.

Autoregressive (AR) and Markov models are of particular interest in restoration of lost speech samples. Esquef et al. [2] recently proposed a time-reversed excitation substitution algorithm with a multi-rate post-processing module for audio gap restoration. Rødbro et al proposed a packet loss concealment using hidden Markov models.

Sinusoidal models, and their extended version HNM, have been applied to speech gap restoration. In [6] the excitation signal is modeled using a sinusoidal model and the LP parameters are repeated for each frame. Rødbro et al. proposed a linear interpolation technique for estimation of the sinusoidal model parameters of missing frames [7].

In this paper we propose a LP-HNM model of speech where the spectral envelope is modeled using a LSF representation of a linear prediction (LP) model and the excitation is modeled with a HNM, whose parameters are the harmonic frequencies, harmonic amplitudes, harmonicities (voicing levels) and phase. The advantage of using LP-HNMs is that the time-varying contours of the formants and the harmonic energies of the signal are tracked and then interpolated across the speech gap to produce a synthesized speech segment that fits the speech characteristics on both sides of the gap.

Furthermore, a codebook-mapping technique is used as a post-processing module for fitting the interpolated parameters to a pre-trained speech model. Codebook-mapping technique has found applications in bandwidth extension [5] and noise reduction [8]. Interpolation of LSF values of speech sometimes results in unusually sharp poles giving rise to tonal artefacts. The conventional method of mitigating such effects is through damping of the poles of LP model at the cost of smearing the signal spectrum [9]. Application of the codebook-mapping technique mitigates the effects of tonal artefacts without an undesirable broadening of the poles' bandwidths. We compare gap restoration methods with the International Telecommunications Union's (ITU) standard for waveform substitution of lost speech signals.

2. LP-HNM MODEL OF SPEECH

In frequency domain the LP-HNM model of the speech, $X(f)$ may be expressed as:

$$X(f) = C \cdot E(f) / L(f) \tag{1}$$

where $E(f)$ is the excitation, $C / L(f)$ is the LP model of the spectral envelope and f is the frequency variable. C is the time-varying gain and $L(f)$ can be modeled with LSF parameters Q. The excitation signal is modeled using a HNM with three parameters: amplitude A_k, harmonicity V_k and harmonic frequency f_k where k is the harmonic index. Harmonicity is a real-valued measure between 0 and 1; it represents the voicing degree of each harmonic sub-band. In the frequency domain, the synthesis Equation for the excitation signal is:

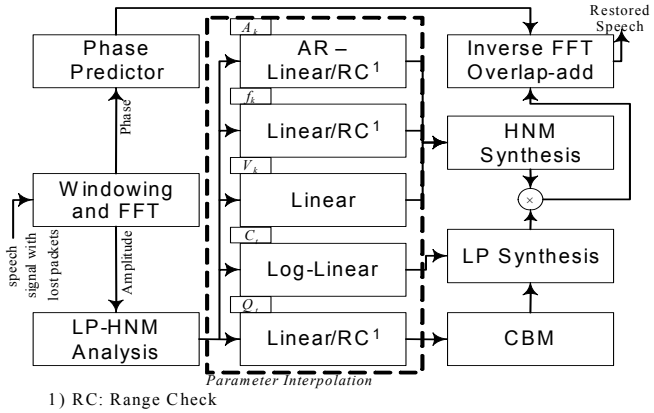

Figure 1. A block diagram of LP-HNM+CBM interpolation system.

$$E_{HNM}(f,t)=\sum_{k=1}^{N} A_{kt}\left[V_{kt}G\left(f-f_{kt}\right)+\left(1-V_{kt}\right)R(f-f_{kt})\right] \qquad (2)$$

where $G(f)$ models the shape of each harmonic which may be set to a delta function or more realistically to a Gaussian-shaped spectrum and $R(f)$ is the noise component of the excitation. The magnitude spectrum of $R(f)$ has a Rayleigh distribution.

3. INTERPOLATION METHOD

Figure 1 shows the proposed interpolation method where speech interpolation is transformed into the interpolation of the LP-HNM frequency-time tracks. The interpolation of each LP-HNM track is achieved using a combination of two methods:

(1) a simple linear interpolation of the mean values across the gap,

(2) a combination of linear and autoregressive interpolation methods.

Assume that T_G consecutive frames of speech are missing where each speech frame has W samples including S new (non-overlapping) samples; and let T_A and T_B be the number of available speech frames after and before the speech gap respectively. Our goal is to estimate the LP-HNM parameters of T_G missing frames using the T_B frames before and T_A frames after the gap. This is shown in Figure 2. It is assumed that the interpolation delay is less than the maximum acceptable system delay.

3.1 Linear Interpolation

Assuming $T_B = T_A = 1$, i.e. only one speech frame is available before and after the gap, the LP-HNM parameters can be linearly interpolated. We define the general linear interpolating function as

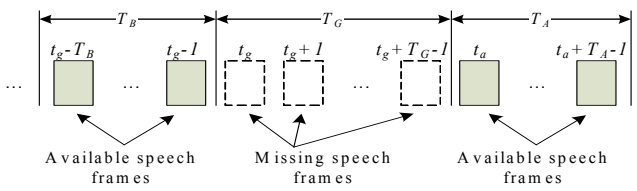

Figure 2. Illustration of a gap of T_G missing speech frames.

$$I_L\left(x_1,x_2,T_G,t\right)=\frac{tx_2+\left(T_G-t+1\right)x_1}{T_G+1} \qquad 1\le t\le T_G \qquad (3)$$

where x_1, x_2 are the known values at the two ends of the gap. Each LP-HNM parameter track, A_{kt}, V_{kt}, f_{kt}, Q_{it} and C_t, can be linearly interpolated using equation (3). Note the subscripts k, i and t represent the harmonic, LSF and frame indices respectively.

The linear interpolation method joins the HNM parameters of speech across the gap with a straight line. Preliminary experiments show that the quality of the interpolated speech is sensitive to estimation error of the excitation harmonic amplitudes and LSF values. Similar results are reported in [8] and [5]. Further experiments show that classical high-order polynomial interpolators also result in artefacts in the output.

3.2 LP-model Interpolation

Note that in this paper we use the terms LP model and autoregressive (AR) model interchangeably. The zero excitation response of a stable LP model, with non-zero initial conditions, decays with time towards zero. The proposed interpolation method exploits this fact in order to obtain an estimate of the parameter sequence which has a smooth transition at each side of the gap and is modeled by the mean values of the LP-HNM parameters in the middle.

Assume the values of the time series x_t are missing from the time instance t_g to t_a-1. One solution would be the least squared error autoregressive (LSAR) interpolator which incorporates information from both sides of the gap simultaneously [1]. However, LSAR assumes that the signals on both sides of the gap are from a stationary process. Furthermore, a large number of samples are required for a reliable estimate of the LSAR models of the time series before and after the gap.

In [9], low-order (4th order) AR models were used to model temporal variations of speech spectrum. Here we use two low order AR models for estimation of the zero-mean trend of the time-series parameter from each side of the gap. The predicted values from each side are overlap-added as shown in Figure 3. The mean value of the time series is estimated by linear interpolation of the mean

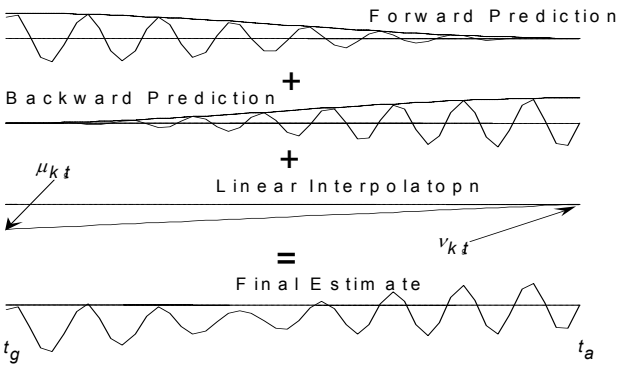

Figure 3. The mean-subtracted time-series is linearly predicted from both sides, weighted-averaged and added to the linearly interpolated mean.

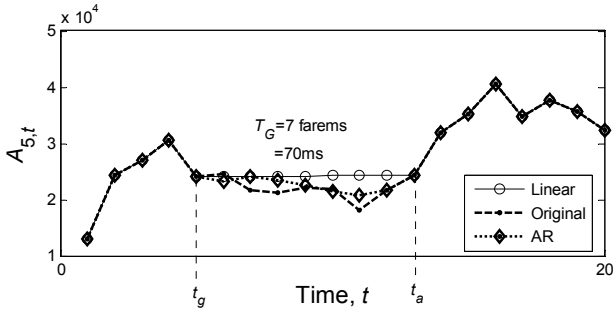

Figure 4 AR interpolation of the 5th harmonic of a sample signal.

values of both sides of the gap. Let the number of available frames at each side of the gap $T_A = T_B \geq 3$. The missing parameter values are estimated as:

$$x_t^{(AR)} = W_t \sum_{i=1}^{P} a_{t_g}^{(i)} \left(x_{t-i} - \mu_{t_g}^{(x)} \right) + \left(1 - W_t\right) \sum_{i=1}^{P} b_{t_a}^{(i)} \left(x_{t+i} - \nu_{t_a}^{(x)} \right) + I_L \left(\mu_{t_g}^{(x)}, \nu_{t_a}^{(x)}, T_G, t - t_g + 1 \right) \quad (4)$$

where P is the order of the LP model and

$$\mu_{t_g}^{(x)} = \frac{1}{T_B} \sum_{i=1}^{T_B} x_{t_g - i} \quad \text{and} \quad \nu_{t_a}^{(x)} = \frac{1}{T_A} \sum_{i=1}^{T_A} x_{t_a + i - 1} \quad (5)$$

are the means and $a_{t_g}^{(i)}$ and $b_{t_a}^{(i)}$ are the i^{th} LP coefficients of the series $\left[x_{t_g - T_B}, ..., x_{t_g - 1} \right] - \mu_{t_g}^{(x)}$ and $\left[x_{t_a + T_A}, ..., x_{t_a + 1} \right] - \nu_{t_a}^{(x)}$, respectively. The weights, W_t's, are chosen from a half of a Hanning window of length $2T_G$. Figure 2 shows the procedure applied in Equation (4). Figure 4 shows an example of the AR interpolation used for interpolation of the 5th harmonic of a sample signal. The length of the gap, T_G, is rather long (i.e. equal to 70ms).

Table 1. Interpolation and post-processing methods used for LP-HNM parameters

	HNM			Envelope AR	
Parameter	A_k	V_k	f_k	Q	C
Method	AR+ Linear	Linear	Linear	Linear	Log-Linear
Post-processing	Range Check >0	Range Check $\in [0,1]$	None	CBM+ Range Check $\in (0, F_S/2)$	None

Table 1 shows the type of interpolation technique used for each parameter. The range-check ensures the interpolated values are within an admissible range. Out of range values are clipped to the maximum or minimum admissible values. The LP-HNM gain is linearly interpolated in the base-10 logarithmic domain.

3.4 Phase Prediction

The phase estimation method is based on a model that exploits the continuity of the harmonic parts of speech and maintains the randomness of the non-harmonic parts. The equation for phase at harmonics is defined [????] as

$$\Phi\left(f_{kt}\right) = \Phi\left(f_{k,t-1}\right) + \frac{2\pi}{F_s}\left[T_{ip}\left(f_{k,t-1} - f_{k,t}\right) + S f_{k,t} \right] \quad (6)$$

where $\Phi(f_{kt})$ is the phase of the k^{th} harmonic at frame t, T_{ip} is the "in-phase" sample index that is where harmonic f_{kt} and $f_{k,t-1}$ are in phase, S is the shift size and F_s is the sampling frequency, the in-phase sample is chosen to be halfway through the overlap as i.e. $T_{ip} = (W + S)/2$ where W is the window size. Furthermore, a level of randomness needs to be added to the phase for unvoiced (noise) synthesis, for each harmonic sub-band of an arbitrary frame:

$$\Phi(f) = \Phi(f_k) + (f - f_k)a + \frac{\Phi_R(f)}{V_k}\psi(f) \quad \text{for } 0 < |f - f_k| < F_0/2 \quad (7)$$

where $a = -\pi W / F_S$ i the slope of the phase, $\Phi_R(f)$ is a random variable uniformly distributed in the range $[-\pi, \pi]$, $\psi(f)$ is a weighting factor that increases with the frequency distance from the centre of the harmonic. We use the following function:

$$\psi(f) = \sum_{k=1}^{N} \left[\left(\frac{1 - h(f - f_k)}{h(f - f_k)} \right) 0.002 + 0.05 \right] \exp\left(\frac{f - 3000}{3000} \right) \quad (8)$$

were $h(f)$ is a hamming window in the range $[-F_0/2, F_0/2]$.

4. CODEBOOK MAPPING

Codebook-mapping (CBM) is a heuristic technique normally used for partial estimation of a set of parameters, e.g. estimation of the upper band's parameters based on those of the lower bands for bandwidth extension [5], or correction of over-suppressed harmonics in a noise reduction system [8]. Codebook mapping forces a model upon the parameters through the use of pre-trained codebooks.

In estimation of the LSF parameters, denoted as Q's, using linear interpolators, we notice that the resulting spectral envelope may have sharp peaks or sound unnatural. These artefacts can be even more annoying than the effect of the original packet loss. One technique that can be particularly useful is damping the poles of the LP model perhaps proportional to the distance from the two ends of the gap [6]. This would mitigate the problem of perceiving sharp peaks in the spectrum at the cost of a de-shaped spectrum.

We propose the use of the codebook mapping (CBM) technique for improving interpolation results and mitigating the effects of unwanted artefacts. CBM technique fits the estimated values into a pre-trained speech model through use of codebooks.

A codebook is trained on LSF parameters of various speech utterances. The utterances were taken from the wall street journal (WSJ) database of spoken language. Each interpolated LSF vector is then compared to the vectors in the codebook and the K nearest

codewords, $\left[\overline{\mathbf{Q}}_{k_1},\cdots,\overline{\mathbf{Q}}_{k_K}\right]$, are selected according to the Euclidian distance:

$$D_k = \left\|\mathbf{Q}^{(L)} - \overline{\mathbf{Q}}_k\right\| \tag{9}$$

where $\mathbf{Q}^{(L)}$ is a linearly-interpolated LSF vector, $\overline{\mathbf{Q}}_k$ is the k^{th} codeword of the LSF codebook, D_k is the Euclidian distance between the two and k_1, k_2 … k_K are the indices of the nearest codewords to $\mathbf{Q}^{(L)}$. These codewords are weighted averaged where the weights are inversely related to their distances from the original LSF vector. The resulting vector replaces the interpolated LSF vector.

$$\mathbf{Q}^{(CBM)} = \left[\sum_{i=1}^{K}\frac{1}{D_{k_i}}\right]^{-1} \times \sum_{i=1}^{K}\frac{\overline{\mathbf{Q}}_{k_i}}{D_{k_i}} \tag{10}$$

Where the superscript (*CBM*) shows the codebook mapped estimate of the LSF vector.

5. EVALUATION

Three different versions of the proposed algorithm are evaluated and compared to some other alternative methods in this section. Besides parametric LP-HNM interpolation with and without codebook mapping, a different method which interpolates the HNM parameters extracted from the speech spectrum itself (and not the excitation) is also evaluated.

The multirate gap restoration algorithm, introduced in the recent work by Esquef, and Biscainho [2], is chosen for comparison purposes. This algorithm is composed of two modules: i) a *core* module which uses an AR model for each side of the gap and estimates the signal using an estimated excitation signal, and ii) a *multirate* post-processing module, which further enhances the interpolated signal in two low frequency sub-bands. In addition to the complete algorithm as introduced in [2], the performance of core method (i.e. without the multirate post-processing) is also evaluated and compared with the proposed algorithms.

Many PLC algorithms proposed in the literature are compared to the standard ITU-T G.711 PLC algorithm [4]. Even though the G.711 PLC algorithm is based on a different set of assumptions than the proposed algorithm, its performance is evaluated compared with the proposed algorithm as a reference point.

A 2-state Markov model is used to model the frame loss introduced in the speech signal. The probability of a "bad" frame after a "good" frame is p and that of a good frame after a bad frame is q [10]. This model emphasizes the burst errors that might occur in some applications.

5.1. Objective Evaluation Results

After introducing the gaps in the signals, each signal is restored using different algorithms, e.g. ITU G.711 PLC algorithm (**G.711**), multirate gap restoration (**Multirate**), the core AR-based algorithm of [2] (**AR**) and the proposed algorithms (**HNM**, **LP-HNM**, **LP-HNM+CBM**).

The performance of these algorithms is evaluated using Perceptual Evaluation of Speech Quality (PESQ) scores and log spectral distance (LSD) measure. The results are calculated and averaged for 100 sentences randomly selected from WSJ database. The performances of different algorithms in for restoration of the gaps generated by a 2-state Markov model are illustrated in Table 2.

Table 2. Performance of different algorithms for restoration of 2-state Markov generated gaps

	PESQ				LSD			
q	0.85	0.7	0.5	0.4	0.85	0.7	0.5	0.4
p	0.1	0.2	0.3	0.6	0.1	0.2	0.3	0.6
Loss Rate %	11	22	38	60	11	22	38	60
Av. Gap Length	1.18	1.43	2.00	2.50	1.18	1.43	2.00	2.50
HNM	3.15	2.73	2.43	2.12	0.52	0.72	0.85	1.06
LP-HNM	3.15	2.74	2.44	2.13	0.52	0.74	0.86	1.00
LP-HNM+CBM	3.14	2.74	2.48	2.21	0.52	0.70	0.81	0.96
G.711 – A1	3.14	2.59	2.07	1.51	1.59	1.99	2.09	2.19
AR	3.00	2.60	2.25	1.77	0.42	0.56	0.64	0.80
Multirate	2.54	2.07	1.73	1.24	0.58	0.77	0.88	1.11
Distorted	2.76	2.01	1.18	0.44	-	-	-	-

5.2. Subjective Evaluation

A set of 5 utterances are selected randomly from the WSJ database. Three different sets of packet loss patterns are generated, using the 2-state Markov model explained in the previous section, with a fixed loss rate of 40 percent and different average gap lengths of 2, 5 and 7 frames [11]. An experiment similar to ITU-T's Comparison Category Rating (CCR) is conducted [12]. After introduction of the gaps, each signal is restored using the three proposed methods and G.711 method. 10 Listeners were asked to listen to the resulting signals, each played after its G.711's restored counterpart and compare the second utterance to the first one and rate it from -3 to 3 representing a "much worse" and "much better" respectively. The results are summarized in Table 3.

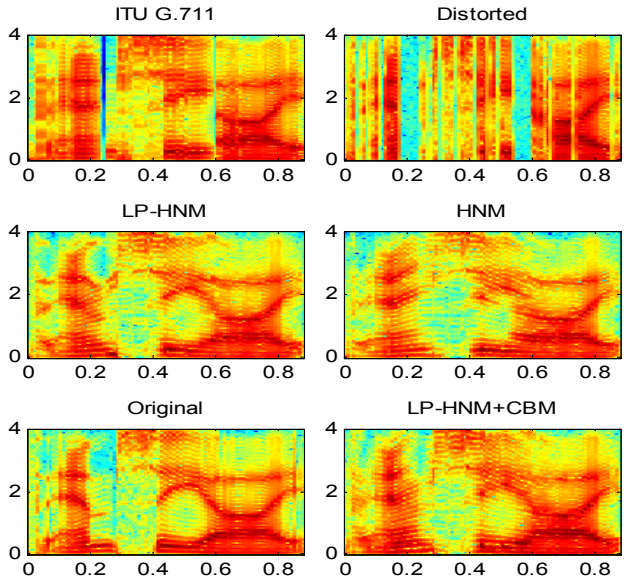

Figure 5. Spectrograms of a sample signal, with introduction of 40% Bernoulli frame loss and the restored versions

5.3. Discussion

Figure 5 shows the spectrograms of a part of a speech signal; it's distorted (with missing samples) and restored versions. It is evident that the restored consonant appeared in the middle of the sample (before 0.4 ms) suffers from different artefacts in different methods. The upper-bands of the speech signal, after restoration with the proposed algorithms, have a higher level of harmonicity compared to interpolation method used in G.711 method. This is due to the more harmonic start of the consonant, available to the algorithms. Freezing effect can be seen throughout the restored gaps of G.711 algorithm which is a known problem of this method. Furthermore, it is observed that the formant trajectories are best recovered using LP-HNM based algorithms.

As mentioned before and generally accepted it is rather difficult to evaluate and compare the performance of speech gap restoration algorithms. Not only each method is designed for a particular application and uses specific resources available, they perform differently in reconstruction of different parts of speech signals. Through exhaustive experiments it was concluded that gap restoration algorithms, in general, are less successful in restoration of vowel-consonant and consonant-vowel transition and even less successful in restoration of vowel-consonant-vowel in which the restored quality is reduced to that of a mumbled speech.

The objective results represented in the previous section shows that the proposed algorithms outperform other algorithms discussed here in most cases. While a very similar output quality is gained in restoration of short gaps, the proposed algorithms are particularly powerful in restoration of longer gaps. The CBM technique proposed in the previous sections results in a level of noise which is believed to be the result of the quantization of the LSF vectors. While at shorter gap lengths this reduces the quality in comparison with some other methods, it makes the algorithm more robust to increases in the gap length particularly for gap length greater than 5 frames as evident in Table 3.

Table 3. Comparative subjective results of proposed methods with a loss rate of 40%.

Restoration Method	HNM			LP-HNM			LP-HNM+CBM		
q	0.5	0.2	0.14	0.5	0.2	0.14	0.5	0.2	0.14
p	0.3	0.13	0.95	0.3	0.13	0.95	0.3	0.13	0.95
Av. Gap Length	2	5	7	2	5	7	2	5	7
Subjective Score	1.64	1.12	0.61	2.36	1.88	0.73	1.68	1.60	1.29

6. CONCLUSION

The problem of restoration of gaps in speech signals was addressed and a solution for reconstruction of the missing parts of the signal resulting in three different algorithms was proposed. It was shown through objective and subjective evaluation tests that the interpolation of the HNM or LP-HNM parameters of speech, used for detailed modeling of the speech envelope and excitation, results in superior output quality. Furthermore, a codebook-mapping technique was employed to make the LP-HNM based algorithm more robust to longer gap lengths. This technique mitigates the problem of tonal artefacts resulting from simple interpolation of LSF parameters at the cost of introduction of some level of quantization noise. We believe that the performance of the proposed LP-HNM+CBM technique can be improved by introducing a gap-length dependency in the CBM technique which will then reduce the level of quantization noise while maintaining the advantage of gap-length robustness of the algorithm. This is being investigated for further improvement of the system.

7. REFERENCES

[1] S. Vaseghi, Advanced Digital Signal Processing and Noise Reduction, 3rd Ed. John Wiley 2006.

[2] P. A. Esquef, L. W. P. Biscainho, "An efficient model-based multirate method for reconstruction of audio signals across long gaps", IEEE Transactions on Audio, Speech and Language Processing, July 2006 Vol. 14, Issue 4, pp. 1391-1400

[3] C.A. Rødbro, M.N. Murthi, S.V. Andersen, S.H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP", IEEE Transactions on Audio, Speech, and Language Processing, vol. PP, Issue 99, 2005 pp. 1 - 15

[4] Appendix I, A High Quality Low-Complexity Algorithm for Packet Loss Concealment With G.711, ITU-T Recommend. G.711, Sept. 1999.

[5] S. Vaseghi, E. Zavarehei, Q. Yan, "Speech bandwidth extension: extrapolations of spectral envelop and harmonicity quality of excitation", ICASSP 2006, vol. 3, pp. III-844 - III-847

[6] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling", in Proc. IEEE Workshop on Speech Coding, Ibaraki, Japan, October 2002, pp. 65-67.

[7] C. A. Rødbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Compressed domain packet loss concealment of sinusoidally coded speech," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc., vol. 1, 2003, pp. 104–107.

[8] E. Zavarehei, S.Vaseghi, Q. Yan, "Weighted codebook mapping for noisy speech enhancement using harmonic-noise model", ICSLP 2006,

[9] E. Zavarehei, S. Vaseghi, Q. Yan, "Interframe modeling of DFT trajectories of speech and noise for speech enhancement using Kalman filters", Speech Communication, *in press*

[10] M.N. Murthi, C.A. Rødbro, S.V. Andersen, S.H. Jensen "Packet loss concealment with natural variations using HMM", ICASSP 2006, vol. 1, pp. I-21- I-24.

[11] [B.P. Milner, A.B. James, "An analysis of packet loss models for distributed speech recognition", Proc. ICSLP 2004, pp. 1549-1552.

[12] Methods for Subjective Determination of Transmission Quality, ITU-T Recommend. P.800, Aug. 1996.