A NEW FEATURE VECTOR FOR HMM-BASED PACKET LOSS CONCEALMENT

L. Koenig[1,2,3], *R. André-Obrecht*[1], *C. Mailhes*[2] *and S. Fabre*[3]

[1] University of Toulouse, IRIT/UPS, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9, France
[2] University of Toulouse, IRIT/INP-ENSEEIHT, 2 rue Camichel, 31071 Toulouse cedex 7, France
[3] Freescale Semiconductor, 134 Avenue du Général Eisenhower - B.P.29, 31023 Toulouse Cedex 1, France
{lionel.koenig, serge.fabre}@freescale.com, corinne.mailhes@enseeiht.fr, obrecht@irit.fr

ABSTRACT

Packet loss due to misrouted or delayed packets in voice over IP leads to huge voice quality degradation. Packet loss concealment algorithms try to enhance the quality of the speech. This paper presents a new packet loss concealment algorithm which relies on one hidden Markov model. For this purpose, we introduce a continuous observation vector well-suited for silence, voiced and unvoiced sounds. We show that having a global HMM is relevant for this application. The proposed system is evaluated using standard PESQ score in a real-world application.

1. INTRODUCTION

In voice over internet protocol (VoIP) networks, voice signal is sent as packets. Due to the different routes used, packets at the receiver may arrive too late for real-time applications, corrupted or may even not arrive. Since in VoIP networks, error-control techniques such as automatic repeat request (ARQ) are not present, the receiver has to tackle the problem of packet loss.
Packet loss concealment (PLC) is an answer to this problem. Three main techniques of PLC can be found in the litterature:

- *Zero insertion* which is simple but obviously not satisfying for the end-user,
- *Packet repetition*: one can choose to reproduce the last frame. Although it sounds better than muting the call, listeners may notice the frame erasure. Better quality can be achieved by using a pitch based waveform replication [3, 5].
- *Model-based repetition*: more advanced methods are trying to fit a model on the speech. When a frame is lost, model parameters are extrapolated and/or interpolated, leading to a recovering of the signal lost part. For example, Gunduzhan proposed a method based on linear prediction [4]. More recently, C.A. Rodbro and al. proposed a PLC based on a hidden Markov model (HMM) [13]. It is based on a semi-hidden Markov model for the speech stream and a minimisation of a mean square error for the concealment.

Although widely used in speech recognition and enhancement, the interest of HMM [11] for PLC has been studied in a very few number of papers. However, results in [13] are promising leading to more natural variations and sounding in the reconstructed speech. Rodbro and al. system relies on a semi-hidden Markov model driven by an unvoiced/voiced estimator. As the feature vector used includes the pitch, the PLC is sensitive to pitch estimation errors like doubling or halving periods.

In this paper we propose a new PLC which has to be independent of the vocoder so that it can be used in any system. We choose to use a unique continuous Markov model for the speech decription to avoid pitch estimation sensitivity.

For that purpose, we propose a new feature vector including an original voicing percentage estimation.

In the section 2 we describe the structure of the proposed HMM-based PLC. Section 3 presents the new continuous feature vector while section 4 focusses on the evaluation of the voicing percentage, which is part of the feature vector. Experimental results of this HMM feature vector are given in section 5. Section 6 concludes this work.

2. HMM-BASED PLC

2.1 Overview

The HMM-based PLC presented first in [9] is directly linked to the vocoder. It assumes that coded frames already include relevant parameters such as spectral envelope, pitch, energy and degree of voicing. Thus, the HMM-based PLC has to produce an estimation of these parameters *before signal synthesis* by the decoder. As a main difference, in the present paper, we propose to introduce a PLC which is independent of the vocoder and can be used in any coding-decoding system, without any *a priori* on the vocoder. Therefore, PLC has to be applied on the decoded speech, *after signal synthesis*.
Moreover, when any PLC is introduced, a choice has to be done:

- either PLC is applied on all received packets, leading to a continuous recovering of the speech without any discontinuity. However, in a perfect packet transmission case, PLC introduces some errors on the reconstructed speech,
- or PLC is applied only on lost packets. This avoids reconstruction errors when the transmission is achieved without any packet loss. However, the produced speech may present some discontinuities which have to be smoothed.

In our work, we choose the second option, leading to the scheme illustrated in Fig. 1. All received frames are analyzed in order to estimate a pre-defined feature vector. When there is a packet loss, the estimation of the missing vector is done through a HMM. In VoIP context, it can be assumed that when considering lost packets, at least one packet corresponding to the speech part located after the missing one is known. This hypothesis has already been done in [13, 9]. Therefore, the estimated vector provided by the HMM takes into account the analysis of frames located before and after the missing speech part. Then this estimated vector, or any related one, is the input of a speech synthesizer. Thanks to the "overlap/add" block, the produced estimated speech is

smoothed in order to reduce discontinuities.

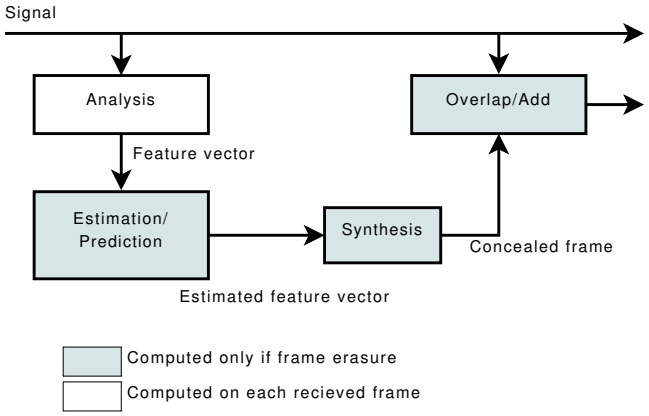


Figure 1: HMM-based PLC architecture

2.2 HMM estimation

For each received frame at time t, a feature vector ϕ_t is computed. This feature vector is composed of relevant parameters which will be detailed in the next section. When some packet loss occurs, let us note L the number of missing packets and $J \geqslant 1$ the number of received packets corresponding to speech part located after the missing part. Then, the missing feature vectors $\phi_{t+k}, k = 0,...,L-1$ are estimated, like in [13]

$$\hat{\phi}_{t+k} = \sum_{n=1}^{P} w_n \mu_{\mathbf{n}} \tag{1}$$

where P is the number of HMM states, μ_n the mean of the n^{th} HMM state, and the weight w_n is the following conditional probability:

$$w_n = \Pr\left(s_{t+k} = n | \phi_1^{t-1}, \phi_{t+L}^{t+L+J-1}\right) \tag{2}$$

with s_{t+k} denoting the random variable representing the state at time $t+k$ and ϕ_i^j the known feature vector from time i to time j.

3. FEATURE VECTOR

In order to avoid discontinuities in the reconstructed speech, the proposed HMM has to be independent of any binary voiced / unvoiced consideration. Therefore, the feature vector proposed should provide a continuous description of the speech signal. Thus we choose a signal representation including the following characteristics:

- A power indicator: the power of the t^{th} frame P_t is computed relatively to the mean variance of the previous frames:
$$P_t = \frac{e_t}{\frac{1}{t}\sum_{j=1}^{t} e_j} \tag{3}$$
with $e_t = \sum_{i=1}^{W} x_t(i)^2$ the energy of the current frame, W the frame size and $x_t(i)$ the i^{th} sample of the frame number t.
- A spectrum description: we describe speech spectral information by 10 Linear Predictive Cepstral Coefficients (LPCC) found by fitting a tenth-order auto-regressive (AR) model to the received speech frames.
- A voicing metric: even if we choose to not distinguish voiced from unvoiced frames, the feature vector has to include some information about the voicing nature of the frame. This voicing indicator has to be "continuous" and not binary to maintain the continuity of the HMM. Therefore, we propose to introduce a parameter defined as the voicing percentage. The next section gives a description and a validation of this new parameter.

4. VOICING PERCENTAGE

4.1 Definition

Voicing percentage $v_{\%}$ is defined as the ratio between the "voicing power" and the overall power of the analyzed speech frame. Voicing power is estimated as the power of the signal frame minus the power of its noise part. To evaluate the spectral part of the noise in the power spectrum density (PSD), we propose to estimate the basis line of the PSD $S(f)$ using a one dimension median filter applied directly on the PSD. The integral of this quantity leads to an estimation of the noise power. The voicing percentage is thus defined as

$$v_{\%} = \frac{\int_0^{0.5} \left(S(\tilde{f}) - \text{median}[S](\tilde{f})\right) \mathrm{d}\tilde{f}}{\int_0^{0.5} S(\tilde{f}) \mathrm{d}\tilde{f}} \tag{4}$$

where $\text{median}[S](\tilde{f})$ denotes the output of a median filter applied to the PSD. Figure 2 illustrates this voicing percentage computation on a voiced frame (a) and an unvoiced one (b). The solid line represents the output of the median filter, while the integral of the solid part of the PSD minus the hatching part corresponds to the numerator of (4). Figure 3 sums up the voicing percentage algorithm.

4.2 Voicing percentage evaluation

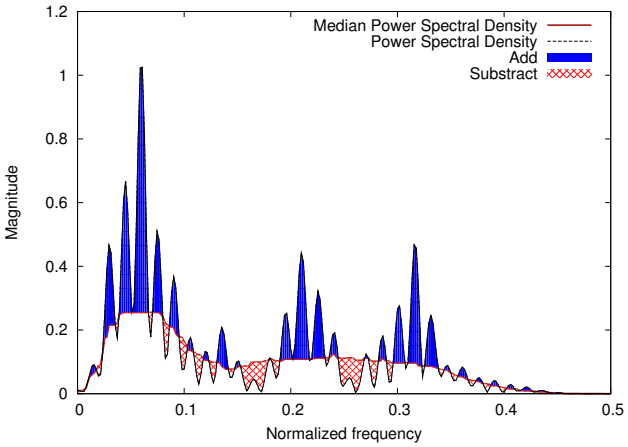
To measure the impact of the voicing percentage on hidden Markov processes, we study it in a classical speech recognition system, more precisely in an acoustic-phonetic decoder. The idea is to see if the introduction of such a parameter in the system will bring or not an improvement of performances.

4.2.1 Baseline decoder

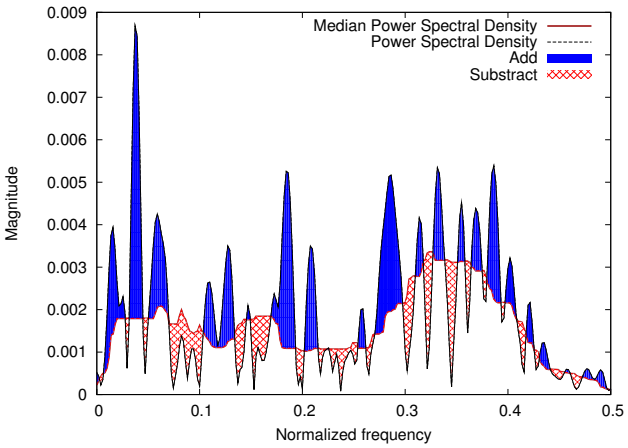
In a first step, a reference acoustic decoder is implemented. This baseline decoder is based on the classical HMM framework. As we use the French corpus BREF80 [7], 35 phones are defined as in [2]. Each phone is modelled with a 3-state-HMM and the observation statistics is assumed to be a Gaussian Mixture Model with thirty-two components. For training, an automatic labelling of the corpus is used [8]. Train and test corpus are described in table 1. Note that no phonetic grammar is introduced.

The HMM uses a 26 component feature vector which includes 12 linear predictive cepstral coefficients, energy and their first order derivative (deltas) like in [12]. A cepstral substraction is performed.

Performances of this decoder which will be related to as a reference one (see table 2) are similar to state of art ones [2].



(a) Voiced frame



(b) Unvoiced frame

Figure 2: Examples of voicing percentage estimation

Part	Subpart	Length
Train	male	4:33:12
	female	5:41:45
	total	10:14:57
Test	male	0:27:46
	female	0:29:58
	total	0:57:44

Table 1: BREF80 Corpus

4.2.2 *Impact of the voicing percentage*

In a second step, in order to assess the interest of the voicing percentage defined in (4) in a HMM system, it has been added to the feature vector of the above implemented acoustic-phonetic decoder. Thus the feature vector includes now linear predictive cepstral coefficients with cepstral substraction, energy, deltas and *voicing percentage*.

The learning process is similar to the one of the baseline system.

Introducing the voicing percentage into the feature vector increases the performances of the phonetic recognition: the

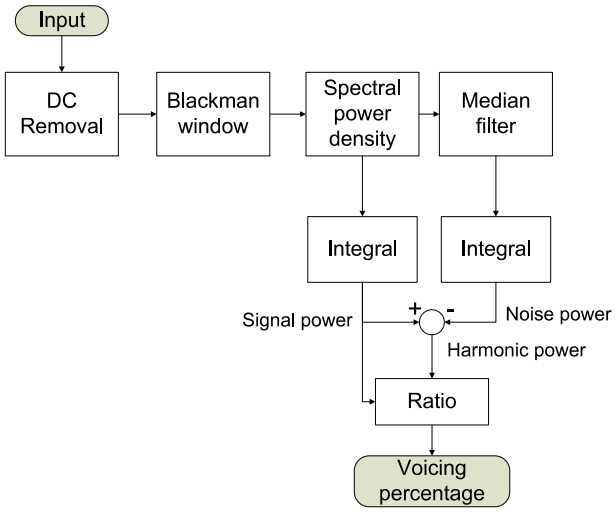


Figure 3: Voicing percentage algorithm

accuracy is about 60.4% while the phone correct rate (PCR) is around 68.4%. These results show the compatibility between non-homogeneous observations and reinforce the use of the voicing percentage in an hidden Markov model.

Model	Accuracy	PCR
Baseline decoder LPCEPSTRA_E_D_Z	59.92%	67.92%
Proposed decoder LPCEPSTRA_E_D_Z + $V_{\%}$	60.39%	68.42%

Table 2: Phonetic speech recognition rates

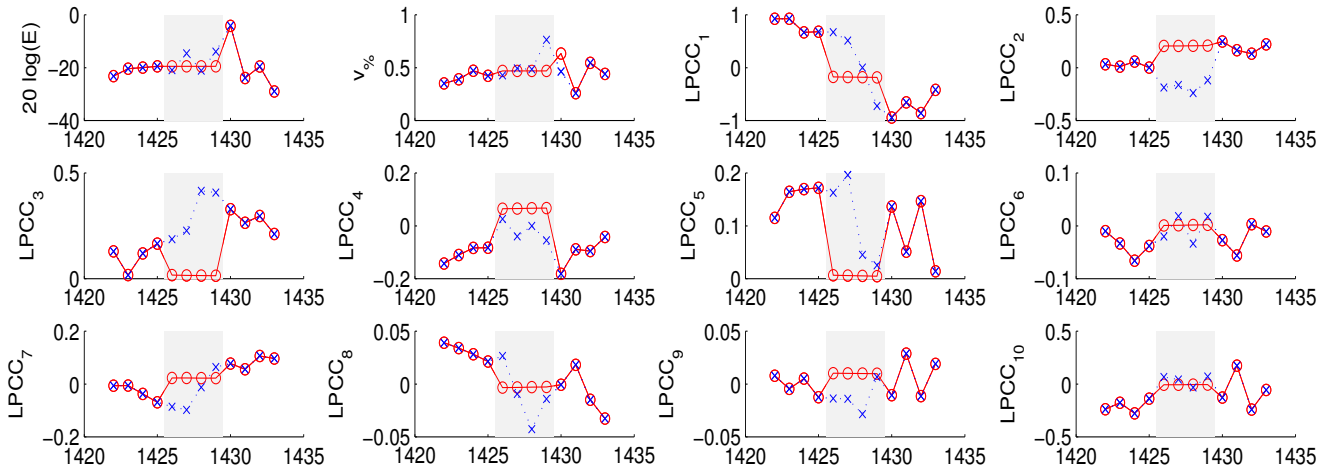
5. EVALUATION - RESULTS

In order to evaluate the interest of the proposed HMM-based predictor as a PLC, two approaches are used.

First, a comparison between the real feature vectors $\phi_{t+k}, k = 0, ..., L-1$ and their corresponding prediction $\hat{\phi}_{t+k}$ is presented. This comparison is made in terms of euclidian distance which is known to be relevant for LPCC [12].

Second, since the final product of a PLC is to reconstruct speech when packets are missing, one has to evaluate the quality of the reconstructed speech when using such a predictor.

For these two approaches, speech signals are extracted from OGI Multilingual Telephonic Speech (OGI MLTS) corpus [10]. Each packet represents $10ms$ of speech signal sampled at $8kHz$. Random loss of L packets ($L \leq 10$) is performed. Due to VoIP architecture, $t-1$ packets before the missing part and J packets after the missing part are assumed to be available. The proposed HMM uses 256 states with one probability density function per state. It was initialized and trained on the English part OGI Multilingual Telephonic Speech corpus using the HTK toolbox.



In this case $L = 4$ and $J = 3$. In dotted line 'x' markers the parameter without loss and in solid line 'o' markers, the estimated parameter in case of packet loss

Figure 4: Estimated vector in a case of a frame loss

5.1 Prediction evaluation

In case of packet loss, the feature vector is predicted using equation (1). Figure 4 presents the evolution of both real vector components (dotted line with o) and their corresponding prediction (solid line with x). During this packet loss, the euclidian distance on LPCC varies from 0.79 to 5.61 which corresponds to acceptable values.

However, it is more valuable to measure directly the quality of the reconstructed speech rather than any distance on any predicted parameter vector. Therefore, in a second step, we introduce a speech synthesizer in order to evaluate speech quality during packet loss.

5.2 Implementation - Speech synthesizer

To assess the quality of the estimated vector, our estimator is coupled with a simple speech synthesizer. The idea here is not to focus on a synthesizer problem but rather to use a well-known classical speech synthesizer [4]. Since the estimated vector $\hat{\phi}_{t+k}$ is based on linear prediction, the use of a linear predictive synthesizer is well-suited. Therefore, the synthesizer used to evaluate speech quality in our study is based on AR coefficients and is presented in Fig. 5.

A 10^{th} order linear filter is matched to the last received frame and used to extract the linear predictive residual signal from the previous frame. This signal is periodized using the pitch of the previous frame. This periodic excitation signal is then filtered through a synthesis filter using the estimated vector produced by the HMM-based estimator as coefficients.

The evaluation of the quality of estimated speech is done with the Perceptual Evaluation of Speech Quality (PESQ) [6] indicator. Table 3 shows PESQ score of the proposed algorithm compared to the PESQ scores obtained with silence insertion PLC or with frame periodization (G711). Losses are generated using a Bellcore model [14, 1] developped by the International Union of Telecommunication.

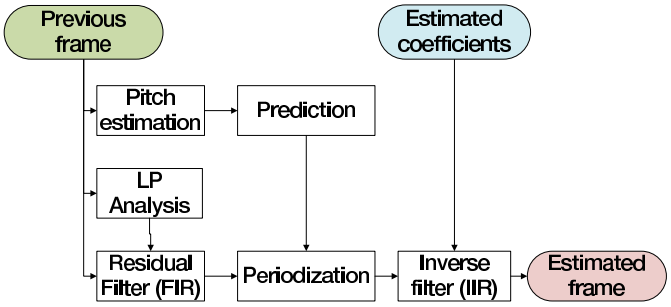


Figure 5: Synthesizer architecture

Corpus	Loss rate	Silence insertion	G711 *Appendix 1*	HMM
OGI MLTS	1 %	3.84	4.07	3.87
	5 %	2.86	3.38	2.91
	10 %	2.24	2.92	2.30

Table 3: PESQ Score

In a loss rate context of 1 to 10%, which corresponds to classical values, the proposed PLC leads to a quality between the silence insertion and G711. However these results are promising since such a HMM-based PLC provides feature vectors of interest which is not the case of the two other considered PLC. These vectors can be used for other simultaneous applications such as speech recognition.

6. CONCLUSION

In this paper we have presented a packet loss concealment based on one hidden Markov model which does not distinguish voiced frame from unvoiced frame by relying on a continuous feature vector. Moreover, this PLC is independent of the speech coder/decoder since it is applied directly on the speech signal.

Promising results shown by this global continuous hidden Markov model stimulates the use of continuous feature vector combined with HMM in the area of estimation. Performances should be compared with [13]. Inner model parameters such as forward or backward variables might be used by external components to perform online speech recognition.

Further work will investigate the impact of the feature vector choice in term prediction/estimation errors. The influence of the HMM structure will also be studied.

REFERENCES

[1] Bellcore. Proposed model for simulating radio channel burst errors. Technical report, CCIT SG XII, 1992.

[2] J.-L. Gauvain and L. F. Lamel. Speaker-independent phone recognition using BREF. In *Proceedings of DARPA Speech and Natural Language Workshop*, Feb. 1992.

[3] D. Goodman, O. Jaffe, G. Lockhart, and W. Wong. Waveform substitution techniques for recovering missing speech segments in packet voice communications. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 11, pages 105–108, 1986.

[4] E. Gunduzhan and K. Momtahan. Linear prediction based packet loss concealment algorithm for PCM coded speech. *IEEE Trans. on Speech and Audio Processing*, 9(8):778–785, 2001.

[5] ITU Recommandation G.711. Pulse code modulation (PCM) of voice frequencies. ITU Recommendation G.711, ITU Recom., Nov 1988.

[6] ITU-T Study Group 12. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU Recommendation P.862, ITU Recom., Feb 2001.

[7] L. Lamel, J.-L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*, pages 505–508, Genoa, Sept. 1991.

[8] O. Le Blouch and P. Collen. Automatic syllable-based phoneme recognition using ESTER corpus. In *ISCGAV'07: Proceedings of the 7th WSEAS International Conference on Signal Processing, Computational Geometry & Artificial Vision*, pages 81–85, 2007.

[9] M. Murthi, C. Rodbro, S. Andersen, and S. Jensen. Packet Loss Concealment with Natural Variations using HMM. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages I.21–I.24, 2006.

[10] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multilanguage telephone speech corpus. In *Proc. of Int. Conf. on Speech and Language Processing*, pages 895–898, Oct 1992.

[11] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[12] L. Rabiner and J. Biing-Hwang. *Fundamentals of speech recognition*. Prentice hall, 1993.

[13] C. Rodbro, M. Murthi, S. Andersen, and S. Jensen. Hidden Markov model-based packet loss concealment for voice over IP. *IEEE Trans. on Audio, Speech and Language Processing*, 14(5):1609–1623, 2006.

[14] V. K. Varma. Testing speech coders for usage in wireless communications systems. In *Speech Coding for Telecommunications, 1993. Proceedings., IEEE Workshop on*, pages 93–94, Oct. 13–15, 1993.