PACKET LOSS CONCEALMENT BASED ON VQ REPLICAS AND MMSE ESTIMATION APPLIED TO DISTRIBUTED SPEECH RECOGNITION

Antonio M. Peinado, Ángel M. Gómez, Victoria Sánchez, José L. Pérez-Córdoba, Antonio J. Rubio

Dpt. de Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada, Spain
{amp,amgg,victoria,jlpc,rubio}@ugr.es

ABSTRACT

This paper proposes a new packet loss concealment technique based on the inclusion in each packet of a few FEC bits, representing data replicas, combined with a minimum mean square error estimation (MMSE). This technique is developed for an Aurora-2 distributed speech recognition system working over an IP network. In addition to the data representing the transmitted speech frames, each packet includes some FEC bits representing a strongly VQ-quantized version (replicas) of previous and subsequent frames. When a loss burst occurs, the lost frames can be reconstructed from the VQ replicas. In order to mitigate the degradation introduced by the coarse VQ quantization of the replicas, a model-based MMSE estimation is applied. The experimental results show that, under a strongly degraded channel, it is possible to obtain up to 83.31 % of word accuracy with only 4 FEC bits or 88.47 % with 8 FEC bits per packet, when the Aurora mitigation algorithm only obtains 76.98 %.

1. INTRODUCTION

When transmitting speech data over a packet network one of the most common problems found is packet loss. Packet losses introduce audio distortions that cause perceived voice quality degradation in the case of IP telephony and a reduction of performance in other speech-based services such as Distributed Speech Recognition [1]. Many packet loss recovery techniques have been proposed which can be broadly classified into two classes: sender based techniques and receiver based techniques [2]. Among the first ones, we have forward error correction (FEC), where repair information is transmitted so that a lost packet can be recovered from that repair data, and interleaving. Among the second class, we have frame repetition, interpolation or more sophisticated regeneration techniques based on signal models.

In this paper we focus on the FEC approach and propose a technique that with very few overhead bits combined with Hidden Markov Model (HMM) based forward-backward minimum mean squared error estimation (FBMMSE) [4] obtains a significant improvement in the performance of a distributed speech recognition system based on the ETSI standard [1] for adverse IP channel conditions. The FBMMSE technique was originally proposed by us in a wireless channel context, where it effectively mitigated wireless channel errors. In the context of the FEC technique for IP transmission, the FBMMSE estimation will be reformulated and used to mitigate the distortion introduced by the coarse quantization of the redundant information included in the proposed FEC scheme.

Work supported by MEC/FEDER project TEC2004-03829/TCM.

The paper is organized as follows. First, the experimental framework is described. Then, we present a review of MMSE estimation. In section 4 we explain the proposed technique in detail and present experimental results. We conclude with section 5, where we discuss the payload format for the ETSI DSR standard and indicate several solutions to introduce the proposed FEC overhead bits.

2. EXPERIMENTAL FRAMEWORK

The front-end utilized in this work is the one proposed in the ETSI standard [1] and developed by the Aurora working group. This front-end segments the speech signal into overlapped frames of 25 ms and provides a 14-dimension feature vector (per frame) containing 13 Mel Frequency Cepstrum Coefficients (MFCC) ($C(k)$ $(k = 0, \ldots, 12)$) plus log-Energy ($\log E$). These features are grouped into pairs and quantized by means of seven Split Vector Quantizers (SVQ). All codebooks have a 64-center size (6 bits), except the one for MFCC-0 and log-Energy, which has 256 centers (8 bits). The bitstream is generated by grouping frames into pairs (88 bits) that are protected by a 4-bit CRC. The Aurora packet loss mitigation algorithm can be summarized as follows: once a loss burst, containing 2B frames, is detected, the first B frames are substituted by the last received frame before the burst and the last B ones by the first received frame after the burst.

The recognizer is the one provided by Aurora-2 and uses eleven 16-state continuous HMM word models (plus silence and pause, that have 3 and 1 states, respectively) with 6 gaussians per state. The training and testing data are extracted from the Aurora-2 speech database. Training is performed with 8440 clean sentences and test is carried out over set *A* (4004 clean sentences distributed into 4 subsets). The performance of the recognition system will be measured in terms of the Word Accuracy (WAcc). The Wacc values obtained with this system are 99.02 % (without quantization) and 99.04 % (after SVQ quantization).

The transmission channel has been modeled by a Gilbert model [3], which is used to simulate six different channel conditions that are summarized in table 1, where *clp*, *ulp* and d_{av} are the packet loss probability when the previous packet has been already lost, the *a priori* probability of a packet loss and the mean loss burst duration (in number of packets), respectively. We will consider that each packet contains two frames (one frame pair).

# Condition	clp	ulp	d_{av}
1	0.147	0.006	1.172
2	0.330	0.090	1.492
3	0.500	0.286	2.000
4	0.600	0.385	2.500
5	0.700	0.500	3.333
6	0.800	0.550	5.000

Table 1. *Description of the simulated channel conditions.*

3. REVIEW OF MMSE ESTIMATION

In our previous work [4] we showed that the MMSE estimation is a powerful technique to mitigate the errors introduced by a wireless channel. In this section we present a brief review of this technique under a formulation suitable for the considered application. The estimation is performed on a feature pair basis, since this is the encoding unit used in the Aurora standard. After the SVQ quantization [1], each feature pair is represented by a vector $\mathbf{c}$ ($\mathbf{c} \in \{\mathbf{c}^{(i)}; i = 0, \ldots, 2^M - 1\}$) ($M$=6,8 in this work). We consider that, at the back-end, the received vector $\hat{\mathbf{c}}$ can be affected by some type of distortion. We also consider that this distortion has a bursty characteristic affecting $T-1$ frames, corresponding $t = 0$ and $t = T$ to the last and first correctly received vectors before and after an error burst, respectively. The MMSE estimation of the received parameter vector at time t, which considers the previous and subsequent received vectors, is obtained as,

$$\tilde{\mathbf{c}}_t = E[\mathbf{c}_t|\hat{\mathbf{c}}_0, \hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_T] = \sum_{i=0}^{2^M-1} \mathbf{c}^{(i)} \gamma_t(i) \quad (0 < t < T) \tag{1}$$

with

$$\gamma_t(i) = P(\mathbf{c}_t^{(i)}|\hat{\mathbf{c}}_0, \hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_T) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=0}^{2^M-1} \alpha_t(j)\beta_t(j)}$$

$$\alpha_t(i) = P(\mathbf{c}_t^{(i)}|\hat{\mathbf{c}}_0, \hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_t)$$

$$\beta_t(i) = P(\hat{\mathbf{c}}_{t+1}, \ldots, \hat{\mathbf{c}}_T|\mathbf{c}_t^{(i)})$$

where $\alpha_t(i)$ and $\beta_t(i)$ are the forward and backward conditional probabilities, respectively. We have also expressed $\mathbf{c}_t = \mathbf{c}^{(j)}$ as $\mathbf{c}_t^{(j)}$ for notation simplicity. The generation of each quantized feature pair is modeled by an HMM model with transition probabilities $a_{ij} = P(\mathbf{c}_t^{(j)}|\mathbf{c}_{t-1}^{(i)})$ and observation probabilities $b_i(\hat{\mathbf{c}}_t) = P(\hat{\mathbf{c}}_t|\mathbf{c}^{(i)})$. The conditional probabilities can be computed from the following forward and backward recursions,

$$\alpha_t(i) = \left[\sum_{j=0}^{2^M-1} \alpha_{t-1}(j) a_{ji}\right] b_i(\hat{\mathbf{c}}_t)/K_t \quad (t > 0) \tag{2}$$

$$\beta_t(i) = \sum_{j=0}^{2^M-1} a_{ij} b_j(\hat{\mathbf{c}}_{t+1}) \beta_{t+1}(j) \quad (t < T) \tag{3}$$

where K_t is a normalization factor at time t. The following initial conditions are applied to $(t = 0)$ and $(t = T)$,

$$\alpha_0(i) = P_i b_i(\hat{\mathbf{c}}_0)/K_0 \qquad \beta_T(i) = 1 \tag{4}$$

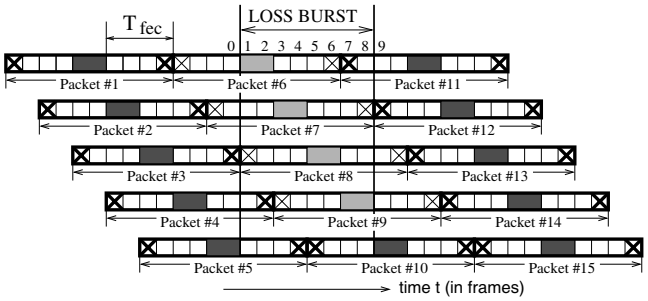

Fig. 1. *Each frame pair (filled with gray) is sent along with a FEC code containing information about the frames (marked with ×).*

where P_i is the *a priori* probability of $\mathbf{c}^{(i)}$.

In our previous work we named this technique as forward-backward MMSE (FBMMSE) estimation in order to remark the use of past and future frames by introducing the forward and backward probabilities and a decoding delay.

4. USE OF VQ REPLICAS AND MMSE ESTIMATION

The problem of applying the FBMMSE estimation described above to the case of a lossy packet channel is that no information is received from the channel during a packet loss period (that is, we do not have vectors $\hat{\mathbf{c}}_t$), what would make the FBMMSE technique useless. A way of solving this problem would be the introduction of information (as FEC bits) about previous and subsequent frames in each packet. Then, there would be some information available about the lost frames during a loss burst. Besides, we would be breaking the burst into shorter bursts and, therefore, increasing the performance. This fact can be easily shown by means of the following experiment. Let us suppose that, along with the feature vectors corresponding to the current frame pair, we also include in the packet exact replicas of the feature vectors corresponding to the frame located T_{fec} frames before the current frame pair and to the frame located T_{fec} frames after the current frame pair. This is depicted in figure 1 for $T_{fec} = 4$. The replicas are marked with the symbol ×. Frames not marked (in white) are not included in the packet. Each packet would then be composed of four frames. The numbering assigned to packets indicates the order in which the packets are sent, while time t is measured in frames. The frame pairs associated to packets lost during a loss burst are indicated in light gray (packets 6, 7, 8 and 9). It can be seen that by using this scheme not all the frames corresponding to lost packets are lost during a burst (frames 2, 4, 5 and 7, marked with bold ×, are recovered). For those frames that are definitively lost (frames 1,3,6 and 8; marked with weak ×), the following mitigation algorithm is applied: for each time $t \leq B$ of a loss burst of length $2B$, the last feature vector received (original or replica) is repeated forwards until a new feature vector is received. For the second half of the burst a similar operation is performed backwards.

The results of this experiment (AURORA+) for $T_{fec} = \{6, 10\}$ are shown and compared with the basic Aurora mitigation algorithm in figure 2. They illustrate the utility of breaking the bursts into shorter ones. This "breaking" idea has been previously and successfully applied for DSR in a different way by performing frame interleaving [5].

However, sending all this additional redundancy increases the bandwidth requirements and, therefore, the loss rate. In order to

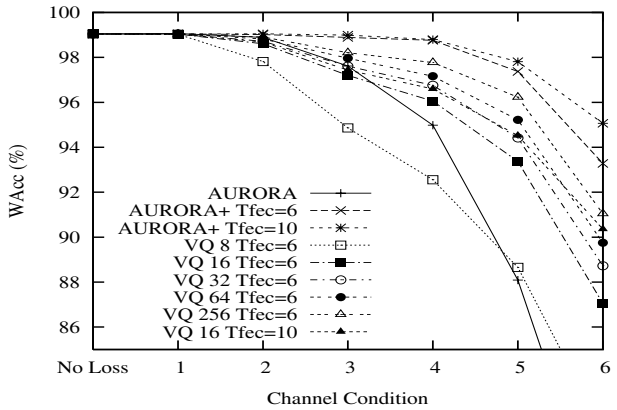

Fig. 2. *Comparison of different mitigation procedures: Aurora, Aurora+ (Tfec=6,10).*

maintain the final bit-rate within a reasonable limit, the repeated feature vectors can be VQ-quantized using a codebook that includes all features (13 MFCCs + logE) with 2^N centroids (N bits). We will use the following distance measure for the codebook design (k-means is used) and in the quantization process,

$$d_W(\mathbf{x}_t, \mathbf{x}_r) = \frac{\sum_{k=1}^{12} (C_t(k) - C_r(k))^2}{\sigma_C^2} + \frac{(C_t(0) - C_r(0))^2}{\sigma_{C\,0}^2} + \frac{(\log E_t - \log E_r)^2}{\sigma_{\log E}^2} \quad (5)$$

where $\mathbf{x}$ represents a 14-dimension feature vector, σ_C^2 is the sum of the MFCC (1-12) variances, and σ_{C0}^2 and $\sigma_{\log E}^2$ are the variances of $C(0)$ and $\log E$, respectively. Each packet should include the following information:

1. 88 bits corresponding to the SVQ-quantized features of the current frame pair.
2. $2 \times N$ bits corresponding to the VQ-replicas.

At the back-end, the VQ replicas can be directly used to improve the recognition. In the case of a feature vector definitively lost, we apply the same mitigation algorithm used in the experiment AURORA+ (that is, repetition of original SVQs or VQ replicas). The results of this strategy are also depicted in figure 2 (experiments VQ) for different VQ codebook sizes (8, 16, 32, 64 and 256 centroids) and $T_{fec} = 6$. It is observed that this technique can provide results similar or better than Aurora, for all channel conditions, using a VQ codebook size of 32 centers (5 bits) or higher, for $T_{fec} = 6$. By increasing the delay T_{fec}, it is possible to improve the performance. As an example, figure 2 also shows that the performance of Aurora is achieved with a VQ codebook of 16 centers and $T_{fec} = 10$ for channel conditions 1, 2 and 3, and is considerably improved for conditions 4 and 5.

Although we have shown that the VQ replicas are useful by themselves, they can be further exploited by performing an FB-MMSE estimation, since we have now information about some of the lost frames. In order to do this, we will divide the received VQ replicas $\mathbf{x}$ into feature pairs $\hat{\mathbf{c}}$. The transition probabilities $a_{ij} = P(\mathbf{c}_t^{(j)}|\mathbf{c}_{t-1}^{(i)})$ of the HMM model are determined from the training data as in [4]. The main difference from the wireless case

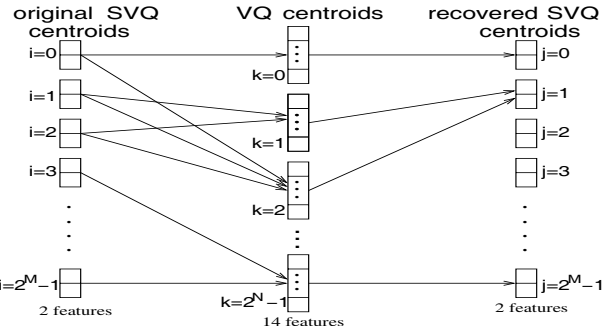

Fig. 3. *Example of the sequence of quantizations applied to the replicas corresponding to one of the SVQ feature pairs.*

treated in [4] resides in the calculation of the observation probabilities $b_i(\hat{\mathbf{c}}_t) = P(\hat{\mathbf{c}}_t|\mathbf{c}^{(i)})$, as the distortion of the received $\hat{\mathbf{c}}_t$ is now due to a strong VQ process and not due to the wireless channel transmission errors. The determination of the observation probabilities $b_i(\hat{\mathbf{c}}_t)$ at each time t ($0 \leq t \leq T$) will be determined, depending on the case, in the following way:

1) In the case that vectors at times $t = 0$ or $t = T$ have been correctly received, the corresponding observation probabilities must be set as,

$$b_i(\hat{\mathbf{c}}_0), b_i(\hat{\mathbf{c}}_T) = \begin{cases} 0 & \mathbf{c}^{(i)} \neq \hat{\mathbf{c}}_0, \mathbf{c}^{(i)} \neq \hat{\mathbf{c}}_T \\ 1 & \mathbf{c}^{(i)} = \hat{\mathbf{c}}_0, \mathbf{c}^{(i)} = \hat{\mathbf{c}}_T \end{cases} \quad (6)$$

2) In the case there is only available a VQ replica at time t ($0 \leq t \leq T$), we will divide the received vector $\mathbf{x}$ into feature pairs that are SVQ quantized again obtaining $\hat{\mathbf{c}}_t$ (as mentioned previously, the SVQ quantization does not involve any reduction of the recognition performance). This process is illustrated in figure 3. Then, we have to determine $b_i(\hat{\mathbf{c}}_t) = P(\hat{\mathbf{c}}_t|\mathbf{c}^{(i)})$. In order to do that, we can see that an original SVQ centroid can correspond to several VQ centroids (depending on the other features different from the considered feature pair). Also, each VQ centroid corresponds to one recovered SVQ centroid, although the contrary can be false (specially when we use a large VQ codebook). This scheme involves the use of a discrete HMM in the FBMMSE estimation, where the observation probabilities $b_i(\hat{\mathbf{c}}_t) = P(\hat{\mathbf{c}}_t|\mathbf{c}^{(i)})$ are obtained from the training database as frequencies of appearance as,

$$b_i(\hat{\mathbf{c}}_t = \mathbf{c}^{(j)}) = \frac{\text{No. recovered symbol } j \text{ given original } i}{\text{No. original symbol } i} \quad (7)$$

It would be also possible to model these observation probabilities by probability density functions (the HMM model would be continuous and the second SVQ process would be unnecessary) where we should select a suitable parametric form for the corresponding *pdf*'s. The discrete version has been selected for simplicity.

3) The third case occurs when there is not any information available at time t ($0 \leq t \leq T$) from the channel. In this case, the forward-backward algorithm progresses without using the observation probabilities, or, equivalently, by considering that the VQ codebook has 1 center (0 bits transmitted). In this case, $b_i(\hat{\mathbf{c}}_t) = 1$ for all i.

Tables 2 and 3 show the word accuracies obtained with the proposed mitigation techniques for several codebook sizes and T_{fec} = $\{6, 10\}$. The best results correspond, as expected, to the FB-MMSE (FB) technique, that obtains excellent results even with

Chan	Mit	Codebook Size				AUR
		4	16	64	256	
1	VQ	98.94	99.03	99.04	99.02	99.06
	FB	99.05	99.05	99.04	99.04	
2	VQ	97.23	98.60	98.81	98.90	98.88
	FB	99.03	99.04	99.04	99.05	
3	VQ	92.67	97.19	97.96	98.20	97.61
	FB	98.14	98.58	98.78	98.76	
4	VQ	89.18	96.05	97.15	97.77	94.98
	FB	96.58	97.63	98.11	98.32	
5	VQ	83.36	93.36	95.22	96.22	88.08
	FB	92.29	95.32	96.00	96.44	
6	VQ	73.94	87.05	89.74	91.04	76.98
	FB	83.31	88.47	90.00	90.98	

Table 2. *Word Accuracy obtained by VQ and FBMMSE (for $T_{fec} = 6$) in comparison with Aurora (AUR) for different channel conditions (Chan).*

Chan	Mit	Codebook Size				AUR
		4	16	64	256	
1	VQ	98.94	99.03	99.04	99.02	99.06
	FB	99.05	99.05	99.04	99.04	
2	VQ	97.12	98.67	98.83	98.91	98.88
	FB	98.99	99.07	99.05	99.07	
3	VQ	92.92	97.43	98.04	98.32	97.61
	FB	98.22	98.71	98.86	98.80	
4	VQ	89.98	96.59	97.43	97.92	94.98
	FB	96.84	98.02	98.35	98.61	
5	VQ	84.67	94.51	95.91	96.72	88.08
	FB	93.01	96.37	96.99	97.60	
6	VQ	77.52	90.35	92.32	93.56	76.98
	FB	85.37	91.73	93.39	94.43	

Table 3. *Word Accuracy obtained by VQ and FBMMSE (for $T_{fec} = 10$) in comparison with Aurora (AUR) for different channel conditions (Chan).*

codebook sizes as low as 4 or 16. The differences between VQ and FBMMSE tend to diminish as the codebook size is increased. This fact is more noticeable for $T_{fec} = 6$ and is the logical consequence of having long gaps in the middle of the loss bursts, in which case, the mitigations tends to the Aurora algorithm in the case of VQ, and to an estimation with uniform distributions for the observation probabilities in the case of FBMMSE (the obtained estimate would only depend on the source model through the transition probabilities a_{ij}).

5. PAYLOAD FORMAT AND IMPLEMENTATION

In this paper, we have proposed a FEC technique that uses data replicas and MMSE estimation to mitigate the effect of packet losses in a DSR system, obtaining excellent results even with very few FEC bits. In this section we will follow the recommendation of reference [6] regarding the payload format for the DSR standard and propose several solutions to introduce the FEC bits that allow the implementation of the proposed technique. Taking into account this recommendation and the constraint of one frame pair per packet, the payload format is the one depicted in figure 4. Packets are aligned into words of 32 bits. As a result, there are 4 free bits that are filled with zeros in [6].

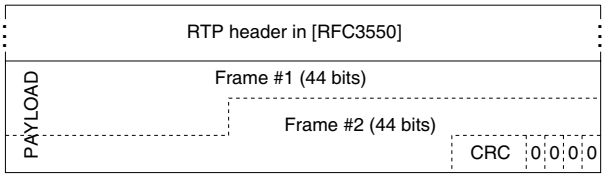

Fig. 4. *Payload format for the DSR standard with one frame pair per packet.*

This payload format suggests us several ways of including the FEC information required for the application of the proposed mitigation methods:
1) We can use the free four bits to introduce two VQ replicas quantized with a 4-center codebook (2 bits/replica). As we can see in tables 2 and 3, in this case it is necessary to apply FBMMSE in order to improve the Aurora results.
2) The system performance can be meaningfully improved if we could reuse the 4 bits devoted to the CRC code to introduce more FEC bits. In this case the replicas are quantized with a 16-center codebook (4 bits/replica). We should use again FBMMSE to improve the Aurora results. However, in this case the VQ technique only provides worse results than Aurora for channel conditions 2 and 3. Since condition 3 corresponds to an average burst duration of 4 frames, a possible solution for mitigation would be a combination of Aurora and VQ. Then, the 2 first and 2 last frames of the burst would be mitigated according to Aurora. The inner $2B - 4$ frames would be mitigated with the VQ technique.
3) Any other increase of FEC would require to include a new 32-bits word in the packet. This case opens a number of new ways for mitigation such as the introduction of 16-bits replicas, the use of 8 bits for the MFCCs and 8 bits for the Energies ($MFCC(0)$ and $\log E$), or even the introduction of four 8-bit replicas. Obviously, these approaches would produce as good or better results than the (best) case of 2 VQ replicas of 8 bits (256 centers) since more information is available.

6. REFERENCES

[1] ETSI ES 201 108 v1.1.2, 2000. Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms. April 2000.

[2] Perkins et al: "A survey of packet loss recovery techniques for streaming audio". *IEEE Network*, vol. 18, pp. 40-48, Sept. 1998.

[3] A.M. Gómez, A.M. Peinado, V. Sánchez, A. Rubio: "A Source Model Mitigation Technique for Distributed Speech Recognition over Lossy Packet Channels". *Proc. of Eurospeech-2003*, pp. 2733-36, Geneve, Sept. 2003.

[4] A.M. Peinado, V. Sanchez, J.L. Perez-Cordoba, A. de la Torre: "HMM-Based Channel Error Mitigation and its Application to Distributed Speech Recognition". *Speech Communication*, vol. 41/4, pp. 549-561, Nov. 2003.

[5] B.P. Milner, A.B. James: "Analysis and Compensation of Packet Loss in Distributed Speech Recognition Using Interleaving". *Proc. of Eurospeech-2003*, pp. 2693-37, Geneve, Sept. 2003.

[6] Q. Xie: RFC 3557 - RTP Payload Format for European Telecommunications Standard ES 201 108 Distributed Speech Recognition Encoding. July 2003.