Interleaving and MMSE Estimation with VQ Replicas for Distributed Speech Recognition over Lossy Packet Networks

Angel M. Gómez, Antonio M. Peinado, Victoria Sánchez,
José L. Carmona, Antonio J. Rubio

Department of Signal Theory, Networking and Communications
University of Granada, Spain

amgg@ugr.es

10.21437/Interspeech.2006-337

Abstract

In this work we evaluate the performance of MMSE estimation with a media-specific FEC based on VQ replicas in comparison with MAP estimation and interleaving, both operating in a DSR system over a loss-prone packet switched network. Both schemes combine a sender-driven with a receiver-based technique and, as we show, clearly outperform the standard Aurora mitigation. However, as independent techniques, interleaving and FEC codes could be jointly applied. Although this would provide better results, a direct combination of FECs and interleaving involves a sum of the delays of both operations. In this work, we introduce a double stream-based strategy that avoids this sum of delays.

Index Terms: distributed speech recognition, loss-prone channels, forward error correction, interleaving, error concealment, maximum a posteriori estimation.

1. Introduction

Packet losses characterize most packet switched networks and can introduce significant limitations to performing Distributed speech recognition (DSR) [1]. Moreover, packet losses tend to appear in bursts and, in DSR, this burst-like nature causes the most negative impact. Thus, DSR has shown to be tolerant to high loss ratios ($\sim$50%) as long as the average burst length is reasonably short (one or two frames) [2].

Media-specific FECs techniques can be especially useful to increase robustness against such losses. These techniques replicate each feature vector in another packet. Indeed, *replicas* can be used not only to recover some lost frames, but also to break bursts of losses into shorter bursts [3]. Since short bursts are better reconstructed, the recognition performance can be improved. However, in order to keep the redundant data into a reasonable size, replicas must be strongly quantized. Exact replacements for lost packets are not obtained and, therefore, an important part of the success of this scheme will depend on the error concealment (EC) technique which manages these degraded replicas.

Alternatively, robustness against bursts can be also increased by applying an *interlaver* prior to transmission. By means of a reordering of the feature vectors, interleaving reduces the perceived burst length at the receiver, improving the recognition performance. As FEC codes, interleaving causes a delay in the transmission but, as advantage, it does not increase the required bandwidth.

In this work, we evaluate a previously proposed FEC-based technique, the MMSE estimation with vector quantized (VQ) replicas [3], in comparison with an interleaver successfully applied to DSR, the optimal delay block interleaver [2, 4]. As we will show, the results individually obtained by these techniques can be improved if both are jointly applied. However, a direct composition of these techniques results in a sum of their delays. In this work, we propose an scheme whereby this increase of delay is avoided.

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

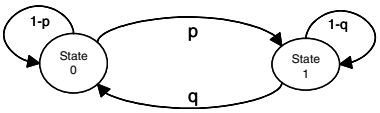

Figure 1: *2-state Markov model. State 0 is error free and state 1 causes frame erasure.*

2. Experimental framework

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group [5]. On the client side, the Aurora DSR front-end segments the speech signal into overlapped frames of 25 ms every 10 ms. Each speech frame is represented by a 14-dimensional feature vector containing 13 MFCCs (including the 0th order one) plus log-Energy. 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). IP packets are generated according to the recommendations of the RTP payload format for DSR [6], where at least two frames (one frame pair) per packet are transmitted in order to avoid too high a network overhead due to headers. Following the RFC recommendations, one frame pair per packet is sent.

The recognizer is the one provided by Aurora [5] and uses eleven 16-state continuous HMM word models, (plus silence and pause, which have 3 and 1 states, respectively), with 3 gaussians per state (except silence, with 6 gaussians per state). The training and testing data are extracted from the Aurora-2 database (connected digits). Training is performed with 8440 clean sentences and testing is carried out over set *A* (4004 clean sentences distributed into 4 subsets).

The channel burstiness exhibited by IP communications is modeled by a 2-state Markov model [7], also known as a Gilbert-Elliot model. Figure 1 depicts this model, where p is the probability of the next packet being lost, provided the previous one has arrived; and q is the probability of the next packet not being lost, given that the previous one was lost. These parameters can be set

in accordance with an average burst length (L_{loss}) and a loss ratio (R_{loss}). The frame numbering included in the RTP header will be used to rearrange the received packets and to detect the frame losses.

3. MMSE estimation with VQ replicas

In a previous paper [3], we introduced a simple media-specific FEC technique that, with very few overhead bits, obtained very good results when combined with a powerful EC algorithm, the forward-backward MMSE estimation (FB-MMSE) [8]. In the proposed FEC scheme, each packet is composed of four frames. Along with the current frame pair, VQ-quantized versions of the feature vectors corresponding to the frames located T_{fec} frames before and after it are included in the packet. These VQ replicas are chosen from a codebook of N bits, which is obtained by a k-means algorithm using the following weighted distance measure:

$$d_W(\mathbf{x}_r, \mathbf{x}_s) = \frac{\sum_{k=1}^{12} (c_r(k) - c_s(k))^2}{\bar{\sigma}_c^2} \quad (1)$$

$$+ \frac{(c_r(0) - c_s(0))^2}{\sigma_{c_0}^2} + \frac{(\log E_r - \log E_s)^2}{\sigma_{\log E}^2} \quad (2)$$

where $\mathbf{x} = (c(0), \ldots, c(12), \log E)$ represents the 14-dimension feature vector, $\bar{\sigma}_c^2$ is the average of MFCCs(1-12) variances, and $\sigma_{c_0}^2$ and $\sigma_{\log E}^2$ are the variances of $c(0)$ and $\log E$, respectively.

These replicas could be directly used but, as we mentioned before, they can be further exploited by applying an FB-MMSE estimation. In order to do so, we will work on a feature pair basis (the encoding unit of the standard). After the SVQ quantization [5], 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.

FB-MMSE estimation is based on an HMM model of speech (further details can be found in [8]). In order to apply it, the transition and observation probabilities of the model, a_{ij} and $b_i(\hat{\mathbf{c}}_t)$ respectively, must be obtained. The transition probabilities a_{ij} can be determined from the training database as in [8]. Regarding the observation probabilities $b_i(\hat{\mathbf{c}}_t) = P(\hat{\mathbf{c}}_t|\mathbf{c}^{(i)})$, we will consider that all feature pairs during a burst $(\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{T-1})$ have been received. These will be determined depending on the type of feature vector considered (received, replica or definitively lost):

- The observation probabilities corresponding to vectors at time $t = 0$ and $t = T$ (assuming they have been received) must be set as,

$$b_i(\hat{\mathbf{c}}_0), b_i(\hat{\mathbf{c}}_T) = \begin{cases} 0 & \mathbf{c}^{(i)} \neq \hat{\mathbf{c}}_0, \mathbf{c}^{(i)} \neq \hat{\mathbf{c}}_T \\ 1 & \mathbf{c}^{(i)} = \hat{\mathbf{c}}_0, \mathbf{c}^{(i)} = \hat{\mathbf{c}}_T \end{cases} \quad (3)$$

- In the case where a VQ replica is available at time t ($0 \leq t \leq T$), it is divided into feature pairs that are again SVQ quantized, obtaining $\mathbf{c}_t^{(j)}$, as Figure 2 illustrates. It is observed that a recovered SVQ centroid can correspond to several VQ centroids, which can also correspond to several original SVQ centroids. Therefore, given an original SVQ centroid $\mathbf{c}^{(i)}$ we can observe several recovered SVQ centroids $\mathbf{c}^{(j)}$ after the double quantization process. It is then possible to determine the observation probabilities from the training database as frequencies of appearance, as follows,

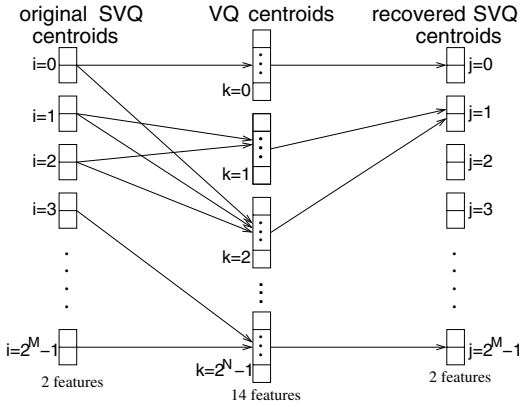

Figure 2: *Example of the sequence of quantizations applied to the replicas corresponding to one of the SVQ feature pairs.*

$$b_i(\hat{\mathbf{c}}_t = \mathbf{c}^{(j)}) = P(\mathbf{c}^{(j)}|\mathbf{c}^{(i)}) = \frac{\text{no. } c^{(j)} \text{ given original } c^{(i)}}{\text{no. original symbol } c^{(i)}} \quad (4)$$

This scheme implicitly involves the use of a discrete HMM in the FB-MMSE estimation. It would also be possible to model these observation probabilities by probability density functions and the second SVQ process would be then unnecessary. However, since SVQ quantization does not involve any reduction in recognition performance [8, 1], the discrete version has been chosen for simplicity.

- Finally, when neither an SVQ vector nor a VQ replica is available at time t ($0 \leq t \leq T$), a degenerated VQ quantization with 0 bits is assumed. Thus, all the original SVQ centroids correspond to only one VQ centroid, the overall mean feature vector, and the observation probabilities are assigned as $b_i(\hat{\mathbf{c}}_t = \mathbf{c}^{(j)}) = 1$, $\forall i, j$. In this case, the forward-backward algorithm mainly progresses guided by the transition probabilities as if the observation probabilities were not used.

This combined technique can significantly improve the robustness against packet losses even with only a few overhead bits. Table 1 shows the results obtained by this scheme using only 4 bits per replica (16 VQ centroids) in comparison with the Aurora standard mitigation (based on the repetition of the nearest received vector). Different delays are considered, corresponding to different values of T_{fec} ($T_{fec} = 6, 12, 20, 30$). It should be noted that by reusing the 8 bits devoted to CRC (only included for compatibility purposes, since IP protocols include their own error protection schemes) and the zero padding bits [6], it is possible to introduce these replicas without any actual bandwidth increase.

4. Interleaving and MAP estimation

An alternative way to break bursts into shorter ones is to permute the order in which complete frames are transmitted. As a consequence, when frames are restored into their original order at the re-

ceiver, consecutive frame erasures are perceived as shorter bursts. To this end, an interleaver can be applied. For an input sequence $\ldots, a_{-2}, a_{-1}, a_0, a_1, a_2, \ldots$ an interleaver can be expressed as a permutation $\pi : \mathcal{Z} \rightarrow \mathcal{Z}$ producing a reordered output sequence $\ldots, b_{-2}, b_{-1}, b_0, b_1, b_2, \ldots$ such that $a_i = b_{\pi(i)}$. Every interleaver has a corresponding deinterleaver that acts on the output of the original interleaver and puts the symbols back into their original order (with a possible time delay δ), that is,

$$\pi^{-1}(\pi(i)) = i + \delta \qquad \forall i. \tag{5}$$

The main advantage of interleaving is that it does not increase bandwidth requirements, but it does have the disadvantage of increasing the delay. There exist different interleavers with different delays, complexities and memory requirements. An interleaver that has been successfully applied to DSR is the optimal delay block interleaver [2, 4, 9]. The block interleaver of degree d operates by re-arranging the transmission order of a $d \times d$ block of input vectors. There are two block interleavers considered optimal in terms of maximizing the spread of bursts for a given degree. They are given by,

$$\pi_1(id + j) = (d - 1 - j)d + i \qquad 0 \leq i, j \leq d - 1, \tag{6}$$

$$\pi_2(id + j) = jd + (d - 1 - i) \qquad 0 \leq i, j \leq d - 1. \tag{7}$$

These two interleavers form an invertible pair, that is, $\pi_1 = \pi_2^{-1}$ and $\pi_2 = \pi_1^{-1}$ and are equivalent to a rotation of the block of feature vectors either 90º clockwise or 90º anti-clockwise (as shown in figure 3). The delay introduced by these interleavers is related to their degree and is equal to $\delta = d(d - 1)$ frames.

Table 2 shows the results obtained by applying optimal delay block interleavers of different degree ($d = 3, 4, 5, 6$). At the receiver, the Aurora standard mitigation is used as EC technique. As can be observed, better results are obtained when the degree of the interleaver, that is, the delay, increases. However, in comparison with table 1, the MMSE estimation based on VQ replicas achieves better results at the only cost of a few overhead bits (that, as we mentioned in section 3, could be introduced without increasing the final bandwidth).

At this point, it can be argued that Aurora standard mitigation is a rather poor EC technique. More advanced techniques have been proposed which exploit statistical information relating to the feature vector stream and provide better results [10]. Maximum a-posteriori (MAP) estimation, for example, replaces the sequence of lost vectors, $\mathbf{X}_m$, by an estimate that maximizes its likelihood conditioned on the received vectors, $\mathbf{X}_o$, and the distribution of the feature vector stream, $P(\mathbf{X}; \mu, \sigma)$. Although it is a coarse approximation, MAP estimation assumes the feature vector stream is Gaussian, so that the MAP estimate reduces to a linear regression entirely described by its mean and variance as [11],

$$\hat{\mathbf{X}}_m = \mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(\mathbf{X}_o - \mu_o) \tag{8}$$

where μ_m and μ_o are the mean vectors of $\mathbf{X}_m$ and $\mathbf{X}_o$ respectively, Σ_{oo} is the auto-covariance matrix of $\mathbf{X}_o$ and Σ_{mo} is the cross-covariance matrix between $\mathbf{X}_m$ and $\mathbf{X}_o$. Due to the inversion of large covariance matrices (Σ_{oo}^{-1}), which imposes too much of a computational overhead, different variations have been proposed to optimize this estimation [10, 2, 4]. In this work we have chosen the fastest one, consisting on a sliding window applied independently over each feature sequence (further details can be found in [10]).

Table 3 shows the word accuracy obtained with MAP estimation using optimal delay block interleavers. As can be observed, this combined application of techniques achieves further improvements in comparison with table 2. These results are even somewhat better than those obtained by MMSE estimation with VQ replicas, but with the additional advantage that they do not involve any bandwidth increase.

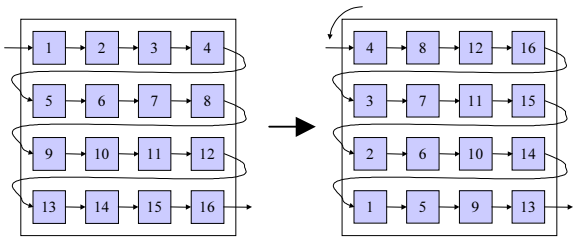

Figure 3: *Illustration of a 4×4 block interleaver. Rotation of 90º anti-clockwise.*

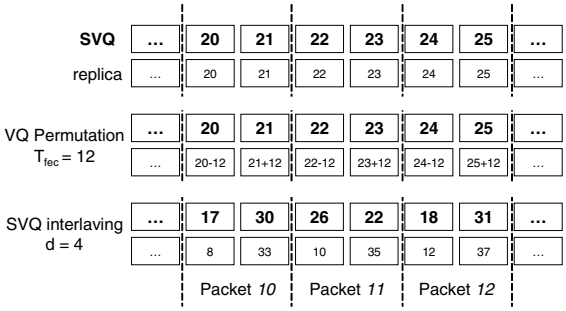

Figure 4: *Example of application of FEC ($T_{fec} = \pm 12$) and interleaving ($d = 4$) in a double stream scheme.*

5. Double-Stream scheme

As independent techniques, there is no reason why FEC codes and interleaving cannot be jointly applied. However, a direct composition of both operations results in a sum of their delays. Thus, if packets are interleaved after FEC codes have been obtained, the delay of the interleaving is added to the delay of the FEC. The same happens when interleaving is applied prior to obtaining the FEC codes.

In this work, we propose an alternative scheme in where this sum of delays is avoided. To this end, feature vectors are grouped in two independent streams. The first stream consists of feature speech vectors coded with SVQ quantization as the standard does, whilst the second stream contains VQ replicas of the first stream. As in section 3, a packet is composed of four vectors: two SVQ vectors from the first stream and two replicas from the second one. Initially, vectors of both streams are ordered by their corresponding time instant. Then, VQ vectors of the second stream are permuted following the proposed FEC scheme, that is, the frames of the current pair of replicas are exchanged with the frame located T_{fec} frames before it and the frame located T_{fec} frames after it. At this point, the resulting packets would be equal to those described in section 3. However, the SVQ vectors of the first stream are now interleaved. In order to do so, the aforementioned optimal delay block interleavers (equations (6) and (7)) can be applied. Figure 4 illustrates the sequence of operations.

At the receiver, the SVQ vectors are restored into their origi-

Condition	*Delay (ms)*				Aurora
R_{loss}, L_{loss}	*60*	*120*	*200*	*300*	
10%, 1	99.00	98.99	99.06	99.03	98.98
20%, 2	98.80	98.79	98.87	98.85	98.08
30%, 4	95.76	97.24	97.72	97.82	90.92
40%, 8	84.65	90.35	93.21	94.58	76.61
50%, 12	72.89	79.76	84.92	87.50	63.27

Table 1: *Word accuracy obtained with MMSE estimation with VQ replicas ($T_{fec} = 6, 12, 20, 30$) in comparison with Aurora.*

Condition	*Delay (ms)*			
R_{loss}, L_{loss}	*60*	*120*	*200*	*300*
10%, 1	99.05	98.98	99.04	99.00
20%, 2	98.79	98.85	98.98	98.95
30%, 4	95.03	96.57	97.75	98.25
40%, 8	83.15	87.07	90.64	93.45
50%, 12	71.06	75.66	80.32	84.46

Table 2: *Word accuracy obtained with block interleaving ($d = 3, 4, 5, 6$) and Aurora standard mitigation.*

Condition	*Delay (ms)*				MAP
R_{loss}, L_{loss}	*60*	*120*	*200*	*300*	estimation
10%, 1	99.02	99.03	99.07	99.01	99.04
20%, 2	98.81	98.86	99.03	98.91	98.31
30%, 4	95.97	97.19	98.03	98.43	93.20
40%, 8	87.08	90.24	92.55	95.04	81.89
50%, 12	76.07	80.65	83.78	88.08	70.19

Table 3: *Word accuracy obtained with MAP estimation with and without block interleaving ($d = 3, 4, 5, 6$).*

Condition	*Delay (ms)*			
R_{loss}, L_{loss}	*60*	*120*	*200*	*300*
10%, 1	99.03	99.03	99.03	99.01
20%, 2	98.92	98.99	99.01	99.00
30%, 4	96.48	98.21	98.72	98.80
40%, 8	86.39	92.19	95.76	97.46
50%, 12	75.77	82.47	88.65	92.21

Table 4: *Word accuracy obtained using a double-stream strategy with block interleaving and MMSE estimation with VQ replicas.*

nal order by means of the corresponding deinterleaver while FEC codes are extracted from packets and used as replicas of lost frames. As in section 3 the MMSE estimation is used to exploit these replicas. Since FEC and interleaving operations are applied over independent streams, this scheme has the advantage of a resulting delay equal to the maximum delay of both operations.

Table 4 shows the results obtained by this scheme for several delays with 4-bit replicas. This double stream-based strategy offers similar or better results than those obtained by interleaving and MAP estimation. Only when the delay is equal to 60 ms, interleaving combined with MAP offers a marginal improvement in the last two conditions (40%, 8 and 50%, 12).

6. Conclusions

In this work we have evaluated the HMM-based MMSE estimation with VQ replicas in comparison with interleaving. As it has been shown, MMSE estimation with VQ replicas performs better than interleaving with a simple mitigation technique, but when interleaving is combined with a statistical-based reconstruction method, the MAP estimation, similar results are obtained, with the advantage of no bandwidth increase.

However, interleaving and VQ replicas could be jointly applied, providing better results than the isolated application of only one of these techniques. Since a direct combination of both operations involves a sum of their delays, we introduce in this work a double stream-based strategy where FEC codes are considered a second virtual stream. Thus, two streams are multiplexed in packets: one with SVQ vectors and the other one with VQ replicas. While VQ replicas are organized following the usual scheme (taking frames at T_{fec} time instants before and after the current frame pair), SVQ vectors are interleaved by means of a block interleaver.

As a result, the proposed strategy achieved the best performance with the advantage of involving a delay equal to the maximum delay of both operations.

7. References

[1] D. Pearce: "Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standard activities for Distributed Speech Recognition Front-ends". *AVIOS 2000: The Speech Applications Conference*, San Jose (USA), May 2000.

[2] B. Milner and A. James, "Robust Speech Recognition over Mobile and IP Networks in Burst-Like Packet Loss", *IEEE Trans. Speech and Audio Processing*, January 2006.

[3] A.M. Peinado, A.M. Gómez, V. Sánchez, J.L. Pérez-Córdoba, A.J. Rubio: "Packet Loss Concealment based on VQ Replicas and MMSE Estimation Applied to Distributed Speech Recognition", in *Proc. ICASSP*, Philadelphia, 2005.

[4] B.P. Milner and A.B. James: "Analysis and Compensation of Packet Loss in Distributed Speech Recognition using Interleaving". in *Proc. Eurospeech*, 2003.

[5] D. Pearce, H. Hirsch: "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions". in *Proc. ICSLP*, vol. 4, pp. 29-32, Beijing, China, October 2000.

[6] "RTP Payload Format for DSR ES 201 108", IETF Audio Video Transport WG, RFC3557, July 2003.

[7] W. Jiang, H. Schulzrinne: "Modeling of Packet Loss and Delay and Their Effect on Real-Time Multimedia Service Quality", in *Proc. NOSSDAV*, June 2000.

[8] A.M. Peinado, V. Sánchez, J.L. Pérez-Córdoba, A. de la Torre: "HMM-Based Channel Error Mitigation and its Application to Distributed Speech Recognition". *Speech Communication*, Vol 41/4, 2003.

[9] A. James, A. Gómez and B. Milner: "A Comparison of Packet Loss Compensation Methods and Interleaving for Speech Recognition in Burst-Like Packet Loss", in *Proc. ICSLP*, 2004.

[10] A. M. Gómez, A. M. Peinado, V. Sánchez, B. P. Milner, A. J. Rubio: "Statistical-based Reconstruction Methods for Speech Recognition in IP Networks", in *Procs. Cost 278 and ITRW workshop*, 2004.

[11] B. R. Ramakrishana: "Reconstruction of Incomplete Spectrograms for Robust Speech Recognition". PhD thesis, Carnegie Mellon University, 2000.