AN INTEGRATED SCHEME FOR ROBUST DISTRIBUTED SPEECH RECOGNITION OVER LOSSY PACKET NETWORKS

Angel M. Gómez, Antonio M. Peinado, Victoria Sánchez, Antonio J. Rubio

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

amgg@ugr.es

ABSTRACT

In this work we present a complete set of techniques devoted to offer robustness against frame losses in distributed speech recognition over packet-switched networks. The proposed scheme is composed of tree techniques, two of them are applied at the sender and the last one in the recognizer itself. On one hand, a media-specific Forward Error Correction (FEC) technique is used to allow the recovery of information within the bursts. On the other hand, a recognizer-based technique well known by its remarkable ability to reduce the effects of long consecutive frame losses during recognition, the weighted Viterbi algorithm (WVA), is used to handle the additional information introduced by FEC codes. Moreover, a double stream strategy whereby interleaving can be applied along with FEC codes without any delay increase, is also applied. The application of interleaving allows to reduce the perceived burst length at the receiver, further improving the recognition performance. As a result, the proposed scheme can provide an acceptable performance even under extremely adverse channel conditions.

Index Terms— Speech recognition, Packet switching, Channel coding, forward error correction, interleaved coding.

1. INTRODUCTION

Nowadays, ubiquitous and pervasive access to information services has become not only desirable but almost necessary. However, the new portable devices, which tend to be smaller and smaller, make this type of access more difficult. Improved user interfaces become necessary and information retrieval through speech recognition arises as solution. The serious constraints that introducing a speech recognition subsystem within those devices would imply, have promoted a novel paradigm known as Distributed Speech Recognition (DSR), where the processing is distributed between the terminal and the network.

Under this approach, the user device extracts and encodes a parametrized representation of speech, which is suitable for recognition. Then, the speech features are transmitted over the network to a remote back-end where recognition is performed. This client-server architecture is very suitable for IP networks, where DSR would allow speech recognition services (and then, an easier access to information) in PDAs, portable thin clients and other handheld devices connected through the Internet. Moreover, powerful centralized recognizers could be shared between multiple users and easily upgraded with new technologies and services.

However, when transmitting real-time data over a packet switched network (such as IP), one of the most common problems encountered is that of packet loss. Since IP networks were designed to offer a best effort service, they are unable to offer a reliable and quality packet delivery. Thus, on congested IP networks, routers will discard packets if their input flow exceeds their output flow for a given data route. Furthermore, irrecoverable errors in the wireless link commonly utilized by handheld devices will result in packet elimination at the back-end. In these scenarios, packet losses tend to appear consecutively and, in speech recognition, this burst-like nature causes the most negative impact. In fact, DSR has shown to be tolerant to very high loss ratios as long as the average burst length is reasonably short (one or two frames) [1].

By means of several mechanisms, it is possible to counteract the effects of bursts of losses, offering an acceptable performance under adverse channel conditions. In this work we propose a double stream strategy of two sender-driven techniques, FEC codes (later called VQ replicas) [2] and frame interleaving [1], which combines with a recognizer-based technique, the weighted Viterbi algorithm (WVA). While interleaving tries to break the bursts into shorter ones, the FEC based technique introduces additional information which will be later managed, along with lost vectors, by the WVA algorithm [3].

The rest of this paper is organized as follows: first, the experimental framework is described; then the media-specific FEC codes based on VQ replicas are explained in section 3. In section 4, the weighted Viterbi recognition with VQ replicas is briefly described and finally, in section 5 the interleaving and the double stream strategy are explained. Experimental results will be provided along the paper.

2. EXPERIMENTAL FRAMEWORK

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group [4]. 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). Following the ETSI standard, 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 organized into a sequence of frame pairs encoded with 88 bits (44 bits per frame) followed by a 4-bit CRC.

The recognizer is the one provided by Aurora [4] 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 speech data has been extracted from clean sentences of the Aurora-2 database (connected digits spoken by American English speakers). Training is performed

Work supported by MEC/FEDER projects TEC2004-03829/TCM and FIT-330503-2006-2.

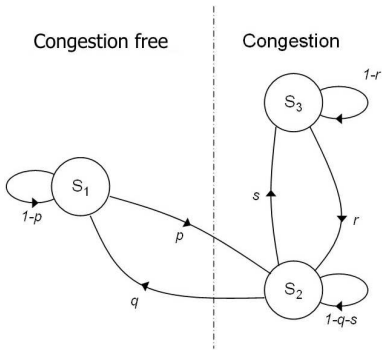

Fig. 1. 3-state Markov model. State 1 and 3 are error free while state 2 causes packet loss.

from a set of 8440 utterances containing a total of 55 male and 55 female adult speakers, and test is carried out over the clean sentences of set A, with 4004 utterances distributed into 4 subsets.

IP packets are generated according to the RTP payload format for DSR [5], which recommends the transmission of two frames (one frame pair) per packet at least, in order to avoid a too high network overhead due to headers. The frame numbering included in the RTP header will be used to rearrange the received packets and to detect the frame losses. The EC technique proposed by the ETSI standard [6] will be used as basic mitigation technique. This technique can be summarized as follows: once a burst, containing $2B$ frames, is detected, the first B frames are replaced by the last frame received before the burst and the last B ones by the first one received after the burst. That is, lost feature vectors are recovered by repeating the nearest received one. The results obtained by this mitigation technique will be taken as our baseline.

The channel burstiness exhibited by packet-switched communications is modeled by a 3-state Markov model [7], with no explicit duration distribution to model burst lengths [8]. Figure 1 illustrates the model topology. In this model four probabilities are considered, namely *p, q, r* and *s*, which characterize two conditions in the network: congested and congestion-free. Packets are steadily received during congestion free periods. This condition corresponds to a low-load state of the network and its period is given by the auto-loop probability of state S_1 $(1-p)$. On the other hand, consecutive bursts of losses appear not excessively distanced in congested condition which corresponds to an overload state of the network. The burst length and the distance between bursts (*ILPL – Inter Loss Period Length*) is given by the auto-loop probabilities of states S_2 and S_3, respectively. The techniques proposed in this work are evaluated under five different channel conditions with packet loss ratios from 10% to 50%, an average burst length of 2, 4, 6, 8 and 10 packets respectively and an average ILPL of 4, 4, 3, 3 and 2, also respectively.

3. VECTOR QUANTIZED REPLICAS

A very simple but effective strategy to counteract the bursty losses caused by the channel is the application of a media specific FEC scheme based on the replication of the feature vectors in packets greatly separated in time [2]. Let us suppose that, along with the feature vectors corresponding to the current frame pair, we also include in the packet replicas of the feature vectors corresponding to the frames located T_{fec} time units before and after the current frame pair. Each packet would then be composed of four frames (figure 2, furthers details are found on [3]). In this way replicas can be used not only to recover some lost frames, but also to break bursts of losses into shorter ones. For example, as can be observed in figure 2, two replicas (marked in gray) are introduced in the middle of the burst, breaking it into two halves.

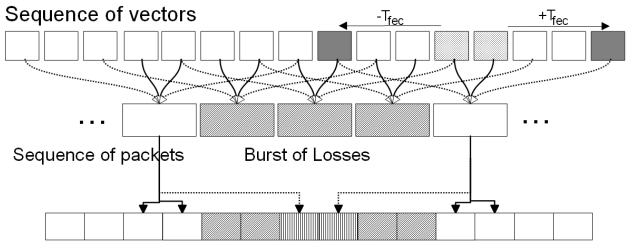

Fig. 2. Each frame pair is sent along with a FEC code containing information about distant frames.

In order to keep the data rate into a reasonable size, the additional vectors are encoded through a vector quantization (VQ) with N bits. The VQ codebook is obtained by applying the k-means algorithm over the 8440 utterances of the training database [1] and using the following weighted distance measure:

$$d_W(\mathbf{x}_r, \mathbf{x}_s) = \frac{\sum_{k=1}^{12} (c_r(k) - c_s(k))^2}{\bar{\sigma}_c^2} + \frac{(c_r(0) - c_s(0))^2}{\sigma_{c_0}^2} + \frac{(\log E_r - \log E_s)^2}{\sigma_{\log E}^2} \quad (1)$$

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

VQ replicas involve an increase both in the delay, since vectors must be buffered prior transmission until replicas in future time instants are provided, and the required bandwidth. Thus, each packet should include, along with the 88 bits corresponding to the SVQ-quantized features of the current frame pair, $2 \times N$ bits corresponding to the VQ-replicas.

Since a frame can be irremediably lost, that is, when neither the packet containing the original vector nor the one containing the replica have been received, this sender-driven technique requires a complementary mitigation technique. We can use the following basic algorithm: for each lost frame at time $t \leq B$ of a loss burst of length $2B$, the last feature vector received (original or replica) is repeated forwards. Analogously, for the second half of the burst ($t > B$) the first received vector is repeated backwards. Table 1 shows the results obtained by applying this scheme in comparison with the Aurora basic mitigation algorithm under the proposed channel conditions. Replicas of 4 and 8 bits have been considered with allowed latencies of 60, 120 and 200 ms. As can be observed, better results are obtained not only by increasing the size of the replicas but also by increasing the allowed delay (T_{fec} parameter). This is because a greater number of distant packets can introduce replicas into the burst.

[1] It would be interesting to test the performance of replicas when VQ codebooks are trained with other databases and under noisy conditions. Future work will address these issues.

Ch.	Aur.	4-bit VQ replicas			8-bit VQ replicas		
		Delay (ms)			*Delay (ms)*		
		60	120	200	60	120	200
1	98.47	98.76	98.73	98.83	98.88	98.93	98.90
2	93.57	94.97	95.93	95.92	95.76	96.59	96.80
3	85.47	89.68	91.75	92.89	91.27	93.75	94.89
4	76.51	81.85	85.25	87.11	84.11	88.45	90.51
5	65.71	69.39	73.85	75.79	72.35	77.71	80.44

Table 1. Results obtained through a direct application of VQ replicas for several allowed sizes and latencies in comparison with Aurora (Aur.).

Ch.	WVA.	4-bit VQ replicas			8-bit VQ replicas		
		Delay (ms)			*Delay (ms)*		
		60	120	200	60	120	200
1	98.60	98.75	98.73	98.80	98.80	98.78	98.88
2	96.30	97.57	97.83	97.86	97.85	98.24	98.34
3	91.97	94.63	95.69	96.11	95.52	96.70	97.34
4	86.28	90.06	91.83	93.24	91.86	93.92	95.24
5	78.87	83.88	86.26	87.65	86.33	89.14	90.96

Table 2. Results obtained through WVR with VQ replicas for several allowed sizes and latencies in comparison with plain WVR.

4. WEIGHTED VITERBI RECOGNITION WITH VQ REPLICAS

Replicas can be further exploited in different ways. One of them is to enhance them using a Forward-backward Minimum mean square estimation (FB-MMSE) as it is proposed in [2, 9, 3]. The other approach is to treat them, along with the very lost frames, at the recognition stage itself. In order to do so, Weighted Viterbi Recognition (WVR) [10, 11], a modification of the Viterbi Algorithm (VA) whereby the confidence in the decoded feature is taken into account, can be applied. As advantage, the replicas degradation (due to the strong quantization) and losses are treated at the recognition stage, where the rich speech source model utilized by the recognizer can be exploited. In this section we will briefly describe the application of WVR with VQ replicas, further details of this technique can be found in [3].

The basic idea of WVR is to incorporate a time-varying reliability γ_t which weights every observation probability in the VA, particularly those referred to lost frames. However, this scheme can be refined by using a reconstruction technique for lost vectors along with a time-varying continuous reliability ($\gamma_{t,k} \in [0,1]$) independently assigned to each feature k (the hypothesis of a diagonal covariance matrix is assumed) [10]. Then, the overall weighted probability can be computed as

$$b_j(\mathbf{x}_t) = \sum_{m=1}^{M} C_{j,m} \prod_{k=1}^{K} \mathcal{N}(x(k); \mu_{j,m}(k), \sigma^2_{j,m}(k))^{\gamma_{k,t}} \quad (2)$$

where M is the number of mixture components, $C_{j,m}$ is a mixture weight and $\mathcal{N}(x(k); \mu_{j,m}(k), \sigma^2_{j,m}(k))$ represents a univariate Gaussian distribution function for the k^{th} feature with mean $\mu_{j,m}(k)$ and variance $\sigma^2_{j,m}(k)$.

An important hurdle to be overcome in WVR with continuous reliability is, indeed, how to determine the reliability function. When Aurora mitigation is applied, an empirically good estimate of the reliability function is based on the normalized auto-correlation of each feature, $\rho_k(\tau)$ [10]. However, $\rho_k(\tau)$ function cannot give coherent reliability values when VQ replicas are used. For this reason, it must be generalized to the normalized cross-covariance between the original lost feature, x, and its estimate, $\tilde{x}$, defined as [3]

$$\bar{C}[x, \tilde{x}] = C_{x\tilde{x}} / \sigma^2_x \quad (3)$$

where σ^2_x is the feature variance and $C_{x\tilde{x}}$ is the cross-covariance between x and $\tilde{x}$, defined as

$$C_{x\tilde{x}} = E[(x - \mu)(\tilde{x} - \tilde{\mu})] \quad (4)$$

where $\mu = E[x]$ and $\tilde{\mu} = E[\tilde{x}]$.

After some lost vectors are recovered by VQ replicas from FEC codes, those definitively lost are replaced by the nearest vector available. Thus, the reconstructed burst is composed of VQ replicas, and repetitions of these and SVQ vectors. The reliabilities for all those features can be obtained as particular cases of equation (3) as follows:

• When SVQ features are repeated, their reliabilities are obtained as,

$$\bar{C}[x_t(k), \tilde{x}_t(k)] = \bar{C}[x_t(k), x_{t+\tau}(k)] \equiv \bar{C}_{SVQ}(\tau; k) \quad (5)$$

• Otherwise, the recovered feature $\tilde{x}_t(k)$ is a repetition of the VQ replica at $t + \tau$ time instant or the replica itself ($\tau = 0$), that is,

$$\tilde{x}_t(k) = VQ[x_{t+\tau}(k)] \quad (6)$$

therefore,

$$\bar{C}[x_t(k), \tilde{x}_t(k)] = \bar{C}[x_t(k), VQ[x_{t+\tau}(k)]] \equiv \bar{C}_{VQ}(\tau; k) \quad (7)$$

where $VQ[]$ represents vector quantization.

Finally, the estimate of the reliability function can be defined as,

$$\gamma_{k,t} = \begin{cases} \sqrt{\bar{C}_{SVQ}(\tau_1; k)} & \text{when } \tilde{x}_t(k) = x_{t+\tau_1}(k) \\ \sqrt{\bar{C}_{VQ}(\tau_2; k)} & \text{when } \tilde{x}_t(k) = VQ[x_{t+\tau_2}(k)] \end{cases} \quad (8)$$

Since the cross-covariance decays relatively quickly, only the normalized cross-covariances for a few τ values must be precalculated (using the training database) and stored, while the remaining ones can be assumed to be zero.

Table 2 shows the results obtained applying this scheme in comparison with a plain (without VQ replicas) system based on WVR with autocorrelation-based time-varying continuous reliability. As can be observed, WVR provides by itself a high robustness against packet losses (in comparison with Aurora mitigation) but this ability can be significantly improved with the additional (although coarse) information provided by the degraded replicas. As before, better results are obtained by increasing the size of the replicas and the allowed delay.

5. INTERLEAVING AND DOUBLE STREAM SCHEME

In order to achieve better results, the proposed FEC scheme based on VQ replicas can be combined with frame interleaving strategies. A frame interleaver permutes the order in which complete frames are transmitted. As a consequence, when frames are restored into their original order at the receiver, consecutive frame erasures are

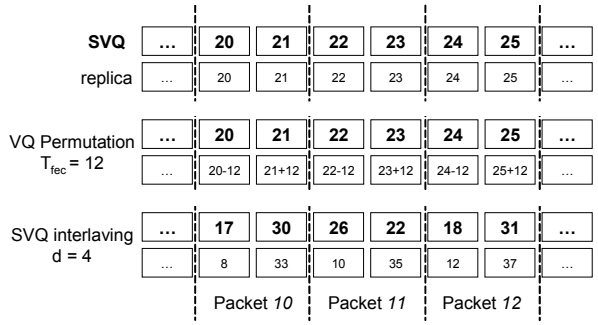

Fig. 3. Example of the application of FEC ($T_{fec} = \pm 12$) and interleaving ($d = 4$) in a double stream strategy.

perceived as shorter bursts. A frame interleaver that has been successfully applied to DSR is the optimal delay block interleaver [1]. It is given by the following invertible pair:

$$\pi_1(id + j) = (d - 1 - j)d + i \quad 0 \leq i, j \leq d - 1, \tag{9}$$

$$\pi_2(id + j) = jd + (d - 1 - i) \quad 0 \leq i, j \leq d - 1. \tag{10}$$

whose delay is related to their degree, d, and is equal to $\delta = d(d-1)$ frames.

However, since both techniques cause a delay, a direct composition of FEC and interleaving schemes results in a sum of their corresponding delays. In this section we propose a double stream strategy whereby this sum is avoided. This scheme operates as follows:

- First, frames provided by the front-end are represented through SVQ centroids as the standard does. However, the same frames are also coded using a VQ codebook of N bits (VQ replicas). The sequence of SVQ vectors will be considered as a primary stream, whilst the sequence of VQ ones will be the secondary stream.
- Primary and secondary streams are processed independently. SVQ vectors are interleaved using the optimal delay block interleaver previously described, while VQ vectors are permuted following the proposed FEC scheme, that is, the frames of each current pair of replicas can be exchanged with the one located T_{fec} frames before it and the one located T_{fec} frames after it. Figure 3 illustrates the sequence of operations.
- Then, packets are built up by taking two SVQ vectors from the first stream and two VQ replicas from the second one. As can be observed, the secondary stream is transmitted within packets as a media-specific FEC.

Since FEC reordering and interleaving are applied over independent streams, this scheme has the advantage of a resulting delay equal to the maximum delay of both operations. As can be observed, if no interleaver is applied on the SVQ vectors the scheme will coincide with that described in section 3. At the receiver, the SVQ vectors are restored into their original order by means of the corresponding deinterleaver while FEC codes are extracted from packets and used as replicas of lost frames.

Table 3 shows the results obtained by means of this strategy when WVR is also applied. As can be observed, this scheme provides the best results presented in this work. Particularly interesting is the case of 4-bit replicas. These replicas can be introduced in the RTP payload for DSR without any actual bandwidth increase [3].

	4-bit VQ replicas			8-bit VQ replicas		
	Delay (ms)			*Delay (ms)*		
Ch.	60	120	200	60	120	200
1	98.87	98.90	98.96	98.875	98.890	98.967
2	98.09	98.53	98.70	98.245	98.665	98.763
3	95.64	97.06	97.76	95.970	97.403	98.022
4	91.51	94.23	95.66	92.272	95.077	96.350
5	85.69	88.62	91.40	87.120	90.160	92.705

Table 3. Results obtained through WVA with VQ replicas and interleaving for several allowed sizes and latencies in comparison with plain WVA.

Therefore, at the only cost of a short delay of 200 ms, it could be possible to maintain the word accuracy above 91% even under extremely adverse conditions (50% of loss ratio with average burst lenghts of 10 packets), while the basic Aurora mitigation only achieves a 65.71% of word accuracy.

6. REFERENCES

[1] B. Milner and A. James, "Robust speech recognition over mobile and IP networks in burst-like packet loss," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, pp. 223–231, 2006.

[2] A.M. Peinado, A.M. Gómez, V. Sánchez, and A.J. Rubio, "Packet loss concealment based on VQ replicas and MMSE estimation applied to Distributed Speech Recognition," *in proceedings of ICASSP*, vol. 1, pp. 329–332, 2005.

[3] A.M. Gómez, A.M. Peinado, V. Sánchez, and A.J. Rubio, "Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels," *IEEE Trans. on Multimedia, In Press*, 2006.

[4] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ICSLP*, vol. 4, pp. 29–32, 2000.

[5] Q.Xie, D.Pearce, S.Balasuriya, Y.Kim, S.H.Maes, and H.Garudari, "RTP payload format for DSR ES 201 108," *IETF Audio Video Transport WG, Internet RFC 3557*, 2002.

[6] *Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI ES 201 108, 2000.

[7] W.Jiang and H.Schulzrinne, "Modeling of packet loss and delay and their effect on real-time multimedia service quality," *Proc.NOSSDAV*, 2000.

[8] A. James and B. Milner, "An analysis of interleavers for robust speech recognition in burst-like packet loss," in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[9] A.M. Gómez, A.M. Peinado, V. Sánchez, J.L. Carmona, and A.J. Rubio, "Interleaving and MMSE estimation with VQ replicas for distributed speech recognition over lossy packet networks," *in proceedings of INTERSPEECH-ICSLP*, 2006.

[10] A.Bernard and A.Alwan, "Joint channel decoding - viterbi recognition for wireless applications," *Proc.Eurospeech*, vol. 4, pp. 2703–2706, 2001.

[11] A. Cardenal-Lopez, L. Docio-Fernandez, and C. Garcia-Mateo, "Soft decoding strategies for distributed speech recognition over ip networks," *Procs. of ICASSP04*, vol. 1, pp. 49–52, 2004.