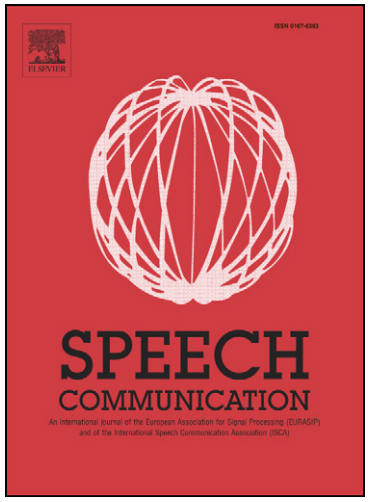
A robust scheme for distributed speech recognition over loss-prone packet channels

Angel M. Gómez, Antonio M. Peinado, Victoria Sánchez, Jose L. Carmona



A robust scheme for distributed speech recognition over loss-prone packet channels

Angel M. Gómez, Antonio M. Peinado, Victoria Sánchez, Jose L. Carmona

PII: S0167-6393(08)00182-9
DOI: 10.1016/j.specom.2008.12.002
Reference: SPECOM 1768

To appear in: *Speech Communication*

Received Date: 8 April 2008
Revised Date: 9 December 2008
Accepted Date: 10 December 2008

Please cite this article as: Gómez, A.M., Peinado, A.M., Sánchez, V., Carmona, J.L., A robust scheme for distributed speech recognition over loss-prone packet channels, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.12.002

A robust scheme for distributed speech recognition over loss-prone packet channels

Angel M. Gómez *, Antonio M. Peinado, Victoria Sánchez and Jose L. Carmona

Dept. Teoría de la Señal, Telemática y Comunicaciones
University of Granada, Facultad de Ciencias, Campus de Fuentenueva S/N
Granada, Spain 18071. Phone: +34 958243271. Fax: +34 958243230.

Abstract

In this paper, we propose a whole recovery scheme designed to improve robustness against packet losses in distributed speech recognition systems. This scheme integrates two sender-driven techniques, namely, media-specific forward error correction (FEC) and frame interleaving, along with a receiver-based error concealment (EC) technique, the Weighted Viterbi algorithm (WVA). Although these techniques have been already tested separately, providing a significant increase of performance in clean acoustic environments, in this paper they are jointly applied and their performance in adverse acoustic conditions is evaluated. In particular, a noisy speech database and the ETSI Advanced Front-end are used, while the dynamic features, which play an important role in adverse acoustic environments, and their confidences for the WVA algorithm are examined. In order to solve the issue of mixing two sender-driven techniques (both causing a delay) whose direct composition causes an increase of the global latency, we propose a double stream scheme which limits the latency to the maximum delay of both techniques. As a result, with very few overhead bits and a very limited delay, the integrated scheme achieves a significant improvement in the performance of a DSR system over a degraded transmission channel, both in clean and noisy acoustic conditions.

Key words: Distributed speech recognition, Media-specific FEC, Interleaving, Weighted Viterbi decoding

* Corresponding author.
Email addresses: amgg@ugr.es (Angel M. Gómez), amp@ugr.es (Antonio M. Peinado), victoria@ugr.es (Victoria Sánchez), maqueda@ugr.es (Jose L. Carmona).

1 Introduction

Distributed speech recognition (DSR) turns out to be a very attractive approach for speech-enabled services over packet-switched networks. As with many other network services, it is based on a client-server architecture where the user device just analyzes, quantizes and sends the speech data to a remote server (back-end) which performs the speech recognition itself. This enables low power/complexity devices to perform speech recognition and allows speech applications to be modeled under a service-oriented approach, in which external information sources can also be linked thanks to the great interoperability of packet-switched networks.

However, when transmitting real-time data over a packet switched network, one of the most common problems encountered is that of packet loss. Packet losses are caused by the inability of IP networks to offer a reliable, high-quality packet delivery service. In fact, in congestion, routers will discard packets if their input flow exceeds their output flow for a given data route. In addition, most portable devices access to the network through a wireless link which is subject to errors. These errors can also be reflected at the back-end as packet losses. In these scenarios, packet losses usually occur in bursts, where multiple consecutive packets are lost (Bolot, 1993).

Consecutive packet losses can have a serious effect on recognition performance (Milner and Semnani, 2000). Although the receiving end could request the retransmission of lost packets, this is not affordable in DSR applications since, as in other real-time applications, it could lead to a worsening of the channel (Xie et al., 2002). Instead, error recovery techniques can be applied to counteract the effects of loss bursts, offering an acceptable performance.

The most extended recovery technique is based on the repetition of the nearest received vector. This simple procedure is included as a mitigation algorithm in the ETSI DSR standards (ETSI-ES201-108, 2000; ETSI-ES202-050, 2002; ETSI-ES202-211, 2003; ETSI-ES202-212, 2005). Other simple error concealment (EC) techniques have also been proposed, such as splicing (Milner and Semnani, 2000) or replacing lost vectors by the mean of training data, but they have been shown ineffective (Endo et al., 2003). Only more complex techniques, such as MMSE and MAP estimation (Gómez et al., 2003, 2004), which use a statistical speech model to estimate replacements, provide better results.

An alternative approach is to deal with packet losses, either fully or partially, inside the recognition engine. In order to do so, reliability information about speech features is taken into account by means of a modified decoding algorithm, as in the weighted Viterbi algorithm (Potamianos and Weerackody,

2001). The main advantage is that a powerful model of the speech, the one present within the recognizer, can be exploited.

On the other hand, the recognition performance can also be improved by shaping the losses into a less damaging distribution. The local stationarity of speech has an important effect on the performance of the mitigation algorithm: short bursts are better reconstructed. Milner and James (2006) have shown that the ETSI DSR error concealment technique makes the recognizer tolerant to channel conditions with very high loss ratios but with short bursts. In clean acoustic conditions, isolated losses can have a negligible effect on recognition performance even at a 50% loss ratio (James and Milner, 2004).

The perceived burst length can be reduced by means of sender-driven techniques such as interleaving (James et al., 2004), but also by forward error correction (FEC) codes, particularly media-specific ones. As we showed in a previous work (Peinado et al., 2005), when packets contain replicas of relatively distant packets (in time), these can be used not only to recover some lost frames, but also to break bursts of losses into shorter bursts. As disadvantage, these sender-driven techniques increase the required bandwidth (FEC) and latency (both, FEC and interleaving). In real-time voice communication, an increase of latency becomes a hurdle. However, in remote speech recognition an immediate response from the recognizer, although desirable, is not strictly required. Thus, an additional delay of a couple of hundred milliseconds may not be too significant to the overall quality of service.

In this work we propose a whole scheme which integrates media-specific FEC codes, frame interleaving and the weighted Viterbi algorithm (WVA) in order to improve speech recognition robustness against packet losses. These techniques have been already tested separately, providing a significant increase of performance in clean environments. In this paper we jointly apply them and, although we are mainly focused on packet losses, their performance in adverse acoustic environments is also tested.

As a starting point we will take a scheme previously proposed in (Gómez et al., 2006) which combines WVA with vector quantized (VQ) replicas (i.e. media-specific FECs). In that work VQ codebooks were trained and tested with the same clean database (although on different training and testing subsets). Here, task-independent VQ replicas (i.e. replicas trained on a different database) have been considered and evaluated while feature confidences used inside the WVA algorithm have been refined for noisy environments, where the dynamic features acquired a special relevance. Then, we tackle the issue of mixing two sender-driven techniques, FEC codes and interleaving (both causing a delay), whose direct composition causes an increase of the global latency. As a result, we propose an integrated scheme which achieves a significant improvement in the performance of a DSR system over a degraded channel with very few

overhead bits and a very limited delay, both under clean and noisy acoustic conditions.

The paper is organized as follows. First, the experimental framework is described in section 2. Section 3 is focused on the modification of the Viterbi algorithm (VA), the calculation of FEC codes and the changes required to treat them in the WVA algorithm. In addition, in this section we also examine the effect of noise over the confidence of delta features. Section 4 is devoted to the interleaving process. In particular, we will provide an intuitive explanation of $(2, t)$ Ramsey interleavers, which have shown very effective in DSR, and test their performance under noisy conditions. Then, in section 5 we tackle the problem of increasing latency and solve it by means of a double stream strategy. In addition, we consider the issue of overlapped replicas and describe a lossy frame interleaver which exploits the redundant nature of replicas. Finally, section 6 summarizes this paper.

2 Experimental Framework

Since it will be useful for the rest of the paper, we will describe here the experimental setup. It is based on the framework provided by the ETSI STQ-Aurora Project Database 3.0. Although the rest of subsets have been used for FEC codebook training (see section 3.3), only the Spanish SDC-Aurora subset has been considered to evaluate our techniques. This database contains 4914 utterances obtained from more than 160 speakers, containing a total of 15924 word realizations. The vocabulary consists of 10 digits between 0 and 9 (zero has one description only) while the recognition task is connected digits. Two different acoustic channels, including recordings from close-talking and hands-free microphones, are considered. Like in other SDC-Aurora databases, recordings are divided into three noisy conditions, namely, quiet, low noisy and high noisy, which provide three different experiments: well-matched (WM), medium-mismatch (MM) and high-mismatch (HM). Each experiment is performed by considering different recording channels and acoustic conditions for training and testing sets (details can be found in (Macho, 2000)). The recognizer is the one included within the database 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 ETSI advanced front-end (ETSI-ES202-050, 2002) is used as feature extractor. This front-end performs the usual cepstral analysis scheme with an additional processing for reducing the influence of background noise. This processing can be roughly divided into three parts: a noise reduction module consisting of a two-stage mel-warped Wiener filter algorithm, a waveform

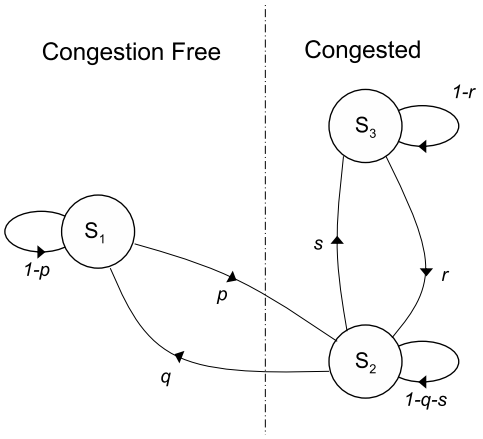


Fig. 1. Three state packet loss model. State 1 and 3 represent no loss while state 2 causes packet loss.

Cond.	R_{loss}	L_{loss}	L_{rec}	p	q	r	s
1	10 %	2	4	0.017	0.125	0.250	0.375
2	20 %	4	4	0.018	0.059	0.250	0.191
3	30 %	6	3	0.020	0.039	0.333	0.128
4	40 %	8	3	0.020	0.023	0.333	0.101
5	50 %	10	2	0.020	0.017	0.500	0.083

Table 1
Loss ratio (R_{loss}), average consecutive loss (L_{loss}) and reception lengths (L_{rec}) together with channel model parameters for tested conditions.

processing module and a blind equalization module. The resulting feature vectors include MFCCs 0-12 plus the log energy. These 14 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 (Xie et al., 2002). This document recommends that, at least, two frames (one frame pair) must be transmitted per packet in order to avoid a too high network overhead due to headers. Nevertheless, the transmission of more frame pairs per packet is not encouraged since longer consecutive frame losses occur when a packet is lost. In this work, only one frame pair is sent per packet.

The channel burstiness exhibited by IP communications is modeled by a 3-state Markov model (Milner and James, 2004). This model is divided into two distinct parts which represent low and high traffic loads on the network (figure 1). Thus, state 1 models non-congestion periods, where packets are steadily received, while states 2 and 3 model congested ones where consecutive bursty losses may appear. As can be observed, burst length is given by the autoloop

probability of state 2 $(1-q-s)$ while the distance between bursts or Inter Loss Period Length (ILPL) (Koodli and Ravikanth, 1999) is obtained through the autoloop probability of state 3 $(1-r)$. ILPL is a relevant measure to take into account when sender-driven techniques are evaluated since the behavior of these techniques not only depends on burst length but also on the time distribution of the receptions. The 3-state Markov model allows to control the average ILPL (L_{rec}) along with the loss ratio (R_{loss}) and the average burst length (L_{loss}).

The recovery methods presented in this paper are tested under the channel conditions listed in Table 1. It can be observed that some conditions show very high amounts of packet loss. However, it should be considered that, at the same loss rate, a fewer number of bursts appear when L_{loss} increases, as the fact that longer bursts imply a greater number of losses. In order to prove the adequacy of the proposed techniques under adverse conditions with long bursts, high amounts of packet loss need to be considered in order to provide a significant number of long burst losses. As a reference, word accuracies for a clean transmission are 80.78%, 87.56% and 93.77% for the HM, MM and WM experiments, respectively.

3 Weighted Viterbi Algorithm with Media-Specific FEC Codes

Weighted Viterbi algorithm (WVA) (Potamianos and Weerackody, 2001) is based on a modification of the Viterbi algorithm (VA) whereby the confidence in the received features can be taken into account. The main idea of WVA is to keep the natural sequence of states in the decoder, conserving the timing information, but applying a weighting coefficient to reduce the effect of unreliable vectors. In order to do so, a time-varying reliability γ_t is incorporated in the VA, obtaining the following state metrics update equation (Yoma et al., 1998):

$$\phi_t(j) = \max_i [\phi_{t-1}(i) a_{ij}][b_j(\mathbf{x}_t)]^{\gamma_t} \tag{1}$$

where $\phi_t(j)$ is the maximum likelihood of observing the feature vector $\mathbf{x}_t$ in state j at time instant t. When the feature vector is fully reliable, γ_t is set to 1 and equation (1) becomes the original state metrics of the VA update equation. On the other hand, when a feature vector is unreliable, γ_t is set to 0. In such a case, the output probability $[b_j(\mathbf{x}_t)]^{\gamma_t}$ becomes 1 for every state and the feature vector has no influence in the selection of the best path.

This technique, usually known as marginalization (Peinado and Segura, 2006), is easy to apply to lossy packet channels: feature vectors can be considered

as fully reliable or completely unreliable depending on whether they have been received or lost. In general, this approach provides good results in the presence of long lost bursts. However, for short bursts, the simple repetition of the nearest received vector clearly outperforms it (Cardenal-Lopez et al., 2006).

3.1 Static feature confidence

The problem in the previous binary approach (full/null reliability) is that it is implicitly assumed that, when a vector is lost, no knowledge can be inferred about its value. However, this is not the case due to the high short-term correlation of the speech signal. Thus, some authors (Bernard and Alwan, 2002; Cardenal-Lopez et al., 2004) have refined the previous scheme by using a time-varying continuous reliability ($\gamma_t \in [0, 1]$) along with a reconstruction technique for lost vectors. Since a speech model is included in the recognizer, complex model-based estimation techniques turn out unnecessary. In their papers, a repetition-based concealment technique is applied and the reliability value is independently assigned to each repeated feature. In order to do so, the hypothesis of a diagonal covariance matrix can be assumed so that the overall weighted probability can be computed as,

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

where M is the number of mixture components, $C_{j,m}$ is the mixture weight and $\mathcal{N}\left(x(k); \mu_{j,m}(k), \sigma^2_{j,m}(k)\right)$ represents a univariate Gaussian distribution function for the k^{th} feature with mean $\mu_{j,m}(k)$ and variance $\sigma^2_{j,m}(k)$. Exponent $\gamma_{k,t}$ is a weighting factor applied to each feature k at time instant t. Thus, the resulting technique is usually known as exponential feature weighting.

Now, the key problem is how to determine the reliability function. It is clear that it is beneficial to decrease the weighting factor $\gamma_{k,t}$ as the feature is consecutively repeated, since the speech signal may have evolved to another sound and the repetition of the received features would no longer be valid. The problem is how to measure this variation. Bernard and Alwan (2002) proposed an empirical reliability function based on the normalized auto-correlation of each feature, $\rho_k(t)$. The validity of this measure has been tested by other authors (James and Milner, 2005; Cardenal-Lopez et al., 2006), concluding that it allowed WVA to achieve an excellent performance with either long or short bursts. Thus, we will assume its use at the moment although, as we will see further, it will be extended in order to be used in our scheme.

3.2 Dynamic feature confidence

An issue frequently forgotten is how to assign a confidence measure to dynamic features. However, dynamic (or delta) features play an important role to obtain a speech representation more robust against acoustic noise. The delta features are obtained as the first and second order derivatives of the static features (MFCCs and Log Energy) usually employing a sliding-window as follows (Furui, 1986; ETSI-ES201-108, 2000):

$$\Delta x_t = \frac{\sum_{w=-W_\Delta}^{W_\Delta} w \cdot x_{t+w}}{\sum_{w=-W_\Delta}^{W_\Delta} w^2} \qquad \text{and} \qquad \Delta\Delta x_t = \frac{\sum_{w=-W_{\Delta\Delta}}^{W_{\Delta\Delta}} w \cdot \Delta x_{t+w}}{\sum_{w=-W_{\Delta\Delta}}^{W_{\Delta\Delta}} w^2} \tag{3}$$

where W_Δ, $W_{\Delta\Delta}$ are respectively the radius of the first and second derivative windows (usually $W_\Delta = 3$, $W_{\Delta\Delta} = 2$). Since dynamic features can be computed at the back end from the static features received, these are not transmitted in DSR systems.

As each temporal derivative is calculated from a window, it is expected that our confidence in it or its reliability is related to the reliability of all the static features within the window. There are some recognizer-based techniques whose mathematical framework allows some kind of estimation of the delta feature reliability (as in (Ion and Haeb-Umbach, 2006)). Otherwise, as in the case of exponential feature weighting, only an heuristic can be proposed as a delta reliability function. James and Milner (2005) proposed and tested several approaches, including the following (from more to less severe):

- Hard decoding. In the computation of the temporal derivatives, if any static feature is lost in the sliding window, the dynamic feature is considered unreliable, that is,

$$\gamma_t^\Delta = \begin{cases} 1 & \text{when } \{x_{t-W_\Delta}, \ldots, x_{t+W_\Delta}\} \text{ are received} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

  where γ_t^Δ is the confidence on the first feature derivative at time t. A similar expression can be written for the confidence $\gamma_t^{\Delta\Delta}$ on the second derivative.
- Product of confidences. The confidence on the temporal derivative is the product of confidences on the static features in the sliding window,

$$\gamma_t^\Delta = \prod_{w=-W_\Delta}^{W_\Delta} \gamma_{t+w} \qquad \text{and} \qquad \gamma_t^{\Delta\Delta} = \prod_{w=-W_{\Delta\Delta}}^{W_{\Delta\Delta}} \gamma_{t+w}^\Delta \tag{5}$$

  where γ_t, is the confidence on the static feature at time t.

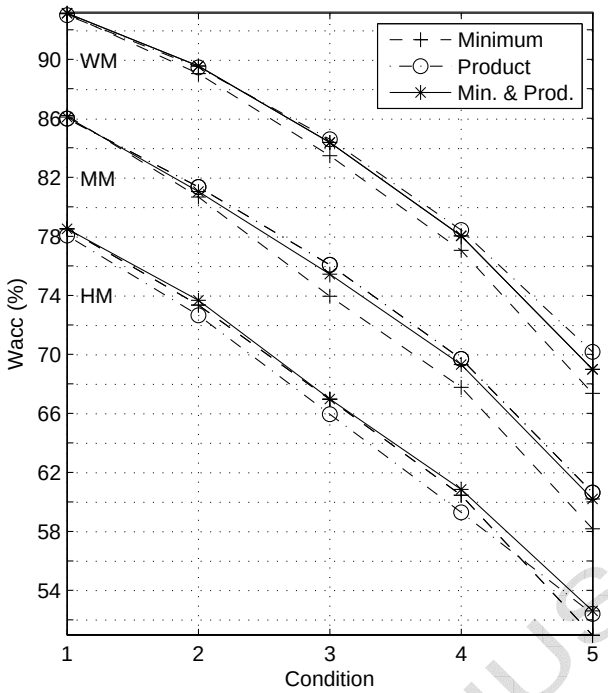


Fig. 2. Performance results obtained with WVA and several delta reliability heuristics for WM, MM and HM experiments under different channel conditions.

- Binary approach. Proposed by Bernard and Alwan (2002), a dynamic feature is considered unreliable ($\gamma_t^{\Delta} = \gamma_t^{\Delta\Delta} = 0$) if the static feature at the same time instant is lost, and fully reliable ($\gamma_t^{\Delta} = \gamma_t^{\Delta\Delta} = 1$) otherwise.
- Regression-based confidences. The confidence on the temporal derivative is obtained through a regression formulation inspired on the equations applied during dynamic features computation (equation (3)). These expressions are given as,

$$\gamma_t^{\Delta} = \frac{\sum_{w=1}^{W_\Delta} w\gamma_{t-w}\gamma_{t+w}}{\sum_{w=1}^{W_\Delta} w} \qquad \text{and} \qquad \gamma_t^{\Delta\Delta} = \frac{\sum_{w=1}^{W_{\Delta\Delta}} w\gamma_{t-w}^{\Delta}\gamma_{t+w}^{\Delta}}{\sum_{w=1}^{W_{\Delta\Delta}} w}. \quad (6)$$

- Minimum of confidences. The confidence on the temporal derivative is the minimum of confidences on the static features in the sliding window,

$$\gamma_t^{\Delta} = \min_{w=-W_\Delta,\ldots,W_\Delta} \{\gamma_{t+w}\} \qquad \text{and} \qquad \gamma_t^{\Delta\Delta} = \min_{w=-W_{\Delta\Delta},\ldots,W_{\Delta\Delta}} \left\{\gamma_{t+w}^{\Delta}\right\}. \quad (7)$$

In terms of computational complexity, the fastest method is the binary approach, as the dynamic feature reliability is directly given by the pattern of losses. Hard decoding, minimun and product of confidences have almost the same complexity while regression-based confidences requires more computations (similar to computing a dynamic feature).

Previous works (James and Milner, 2005; Bernard and Alwan, 2002) have shown no significant differences among the previous methods in acoustically

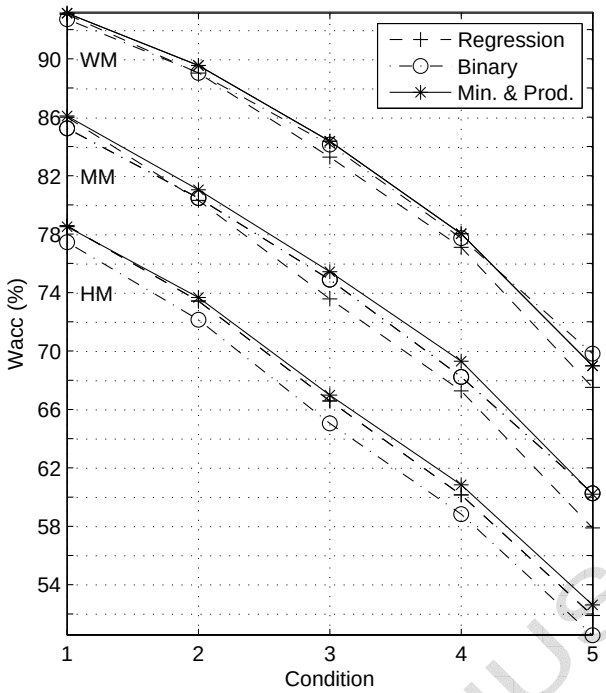


Fig. 3. Comparison between Binary, Regression and our delta reliability heuristic for WM, MM and HM experiments under different channel conditions.

favorable environments (except hard decoding which provides poor results). However, we have found that in noisy conditions, where delta features have a relevant role, the differences become significant. In general, it can be said that restrictive methods perform better in medium mismatch and well-matched conditions whilst less severe approaches achieve better results in high mismatch ones. Figure 2 shows the results obtained considering the product and the minimum of confidences, whilst figure 3 those from the regression method proposed by James and Milner (2005) and from the binary strategy proposed by Bernard and Alwan (2002), both under transmission channel conditions from section 2. As observed, the product of confidences performs better than the minimun (a less restrictive method) in medium mismatch and well-matched conditions, with differences up to 2.46%, but worse in high mismatch ones. The same can be said about the binary strategy which performs slightly better than regression in medium mismatch and well-matched conditions but falls in high mismatch ones (between 1.12% and 1.54%).

A combination of the minimum and product approaches can properly perform in well-matched, medium mismatch and high mismatch conditions. In our proposal the confidence on the first derivative is obtained as the minimum of confidences on the static features but as the product of them for the second

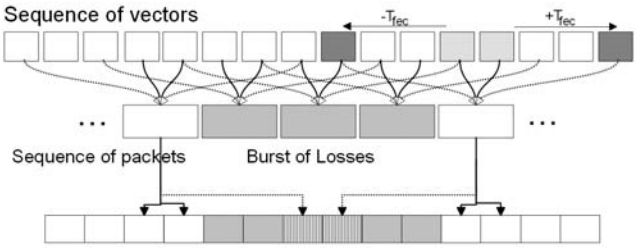


Fig. 4. Each frame pair is sent along with a FEC code containing information about distant frames.

derivative, that is,

$$\gamma_t^{\Delta} = \min_{w=-W_{\Delta},\ldots,W_{\Delta}} \{\gamma_{t+w}\} \qquad \text{and} \qquad \gamma_t^{\Delta\Delta} = \prod_{w=-W_{\Delta\Delta}}^{W_{\Delta\Delta}} \gamma_{t+w}^{\Delta} \tag{8}$$

This approach turns out simpler than regression-based confidences and, in comparison, provides better results (up to 2.31% in MM experiment). Figures 2 and 3 also include the results obtained with this approach (Min. & Prod.). In general, the proposed heuristic does not provide statistically significant differences in well-matched conditions. However, in medium mismatch conditions it approximates or slightly improves the results from restrictive methods (product and binary), whilst in high mismatch conditions, maintains the performance as the less severe ones (minimum and regression).

3.3 Introduction of FEC codes

Weighted Viterbi algorithm is a particularly attractive solution to be used with media-specific FEC. By means of the reliability values it is now possible to inform the recognizer that we have some information about lost frames, the FEC codes. However, our confidence in them is lower, as they have been degraded by a (possibly) strong quantization process.

In our FEC scheme each packet includes the feature vectors corresponding to the current frame pair, as well as replicas (FEC codes) of other vectors. Initially, we can use the same distribution presented in previous papers (Peinado et al., 2005; Gómez et al., 2006), where vectors (replicas) corresponding to the frames located T_{fec} time units before and after the current frame pair are also included within the packet. Figure 4 depicts an example of this scheme. As can be observed, the goal is to insert replicas (marked in light) into bursts, breaking them into shorter bursts.

The replicas are coded with a secondary encoding which requires fewer bits. This keeps the bandwidth increment into a reasonable size and avoids imposing

too much overhead to the network. In our case, each replica, containing the 14 features (13 MFCCs plus log-Energy) is vector quantized (VQ) using a codebook with 2^N codewords (N bits). In our previous works (Peinado et al., 2005; Gómez et al., 2006), VQ codebooks were trained using the k-means algorithm over the speech recognizer training database. Although different sets were used for training and testing, replicas could be over-adapted to the database and, in particular to its vocabulary. In this paper we intentionally check this issue by using different training and testing databases. Thus, VQ codebooks are trained over the Finnish, English, Italian and German SDC-Aurora database subsets, whilst tests are performed over the Spanish subset. In addition, noisy recordings are also considered in both cases.

In our work we are interested in very short replicas of 4 or 8 bits (the reasons will be clear along this paper). However, if such replicas are considered as original vectors, i.e. without any post-processing, a reduction in performance can be observed (Peinado et al., 2005). Therefore these replicas must be treated in a different way by the recognizer. To do so, we can take advantage of the WVA algorithm.

The problem is now that when VQ replicas are used, the autocorrelation-based estimate for reliability mentioned in section 3 is no longer valid. Now, during reconstruction, some lost vectors are recovered through VQ replicas whilst vectors that are definitively lost are replaced by the nearest vector received (either an SVQ vector or a VQ replica) [1]. To tackle the problem of confidence assignment, the normalized auto-correlation function is generalized to the normalized cross-covariance between the original lost feature, $x_t(k)$, and its replacement, $\tilde{x}_t(k)$. Given the aforementioned recovery for lost frames, reliabilities for the replaced features can be obtained as a function of the temporal difference, τ, between the original feature and the repeated one. When SVQ vectors are repeated, the normalized cross-covariance for the repeated feature k, $\bar{C}_{SVQ}(\tau; k)$, is used. When VQ replicas are repeated, the normalized cross-covariance for the repeated and VQ quantized feature k, $\bar{C}_{VQ}(\tau; k)$, is applied instead. Details about the normalized cross-covariance as reliability function can be found in (Gómez et al., 2006).

Table 2 shows the average word accuracies (among WM, MM and HM experiments) obtained by this technique for different codebook sizes and channel conditions (dynamic feature confidences were calculated following the combined Min-Prod strategy described in section 3.1). Results offered by the Aurora standard (repetition) and a plain WVA scheme (without replicas) are

[1] A more elaborated mitigation algorithm could be used, for example, the FB-MMSE estimation with VQ replicas proposed in (Gómez et al., 2006), which can also provide a confidence measure for estimated features (Peinado et al., 2006). However, preliminary results showed no accuracy improvements in comparison with the approach reviewed here.

			4-bit VQ replicas				8-bit VQ replicas			
			Delay (ms)				*Delay (ms)*			
Ch.	Aur	WVA	*60*	*120*	*200*	*300*	*60*	*120*	*200*	*300*
1	85.89	85.97	86.27	86.33	86.24	86.50	86.67	86.62	86.77	86.76
2	78.95	81.47	82.46	82.86	83.28	83.65	84.17	85.00	85.31	85.54
3	71.54	75.44	77.43	77.83	78.83	79.38	80.96	82.25	82.56	83.20
4	63.48	69.43	71.52	73.07	73.33	74.20	76.09	78.41	79.02	79.62
5	54.10	60.62	63.45	65.08	65.43	66.49	68.43	71.85	71.91	73.32

Table 2
Average word accuracies obtained by WVA with 4 and 8-bit VQ replicas for different allowed delays in comparison with Aurora and plain WVA technique.

also included as reference. As can be seen, WVA with VQ replicas provides a significant improvement even with coarse replicas of 4 bits. By increasing the latency we obtain better results. In such a case, replicas are more distant in time so longer bursts can be broken and, therefore, WVA performance improves. On the other hand, as expected, increasing the VQ codebook size from 4 bits per replica to 8 bits also improves the results. However, it must be noted that 4 bit replicas could be inserted within the stream at no bandwidth cost. According to the DSR payload recommendation for RTP protocol (Schulzrinne et al., 1996), 4 bits are devoted to the CRC code while 4 padding bits are filled with zeros to ensure datagram word alignment. Since the 16 bit checksum of the User Datagram Protocol (UDP) (Postel, 1980), along with the codes usually applied at the physical layer and the coherency test performed at the back end, seem sufficient to ensure data integrity, these 8 bits (CRC and padding) could be reused to include two 4-bit replicas.

Finally, as we mentioned before, VQ codebooks have been trained over all the SDC-Aurora databases except the testing one (Spanish SDC-Aurora). We have observed only a slight increase of performance when the same database is used for training and testing (with different sets). These task-adapted FEC codebooks could be also used assuming some mechanism for FEC codebook update prior to transmission. However, we have preferred the first option for the sake generality.

4 Low-latency Frame-level Interleaving

Interleaving is a technique commonly applied at the bit level to disperse the appearance of errors, thus reducing the effect of error bursts. Unfortunately, bit-level interleaving turns out useless when complete packets of information are lost, as in packet-switched networks. In these networks, complete frames can be interleaved instead (Perkins et al., 1998; James and Milner, 2004) by

permuting the order in which these are encapsulated into packets. In such a way, consecutive packet losses are perceived as isolated frame erasures or, at least, as shorter bursts at the receiver. We will refer to these interleavers as frame-level interleavers.

Frame interleavers disperse frame losses relying in the fact that EC is much more effective for short loss bursts. This ability to disperse consecutive losses (or errors) is related to the interleaver *spread*, which is better understood under the point of view of the de-interleaver. Thus, an interleaver has spread (s, t) if (and only if) its de-interleaver has spread (t, s) (Ramsey, 1970), that is,

$$|\pi^{-1}(i) - \pi^{-1}(j)| \geq s \qquad \text{whenever,} \qquad |i - j| < t. \tag{9}$$

where $\pi^{-1}(i)$ represents the original order of the frame received at time instant i. From equation (9) it can be concluded that an interleaver with spread (s, t) will disperse any burst of frame losses with length less than t into isolated frame losses separated by at least $s - 1$ frames (Gómez et al., 2007).

The main drawback of interleaving is the latency involved and, as expected, interest is focused on finding those interleavers which provide the maximum spread causing the shortest possible latency. Thus, very common ones are the minimal latency block interleavers (MLBI), proposed by Andrews et al. (1997), whose latency, given by $2s^2 - 2s$, is minimal among block interleavers of spread (s, s).

MLBI interleavers have been tested by Milner and James (2006, 2003) showing improvements on DSR robustness against bursty losses. These interleavers are easily obtained by a simple rotation of a block of $s \times s$ elements either 90^o clockwise or 90^o anticlockwise. However, they present an important hurdle: the separation the interleaver introduces between two consecutive losses is equal to the length of the bursts it can counteract. In DSR, such a long distance could be excessive, as we previously showed in (Gómez et al., 2007). For example, let us consider the ETSI DSR mitigation algorithm in which lost frames are simply replaced by the nearest received one. The presence of more than one frame between isolated losses will not provide better replacements, since the EC technique can not take advantage of the additional frames. Then, an interleaver with spread $(2, t)$ would offer a similar EC performance than one of spread (s, t) with $s > 2$, but causing a shorter latency. In general, we showed that assuming $s = 2$ is suitable even when applying more advanced EC techniques that make use of two or more received frames to obtain the replacements (Gómez et al., 2007).

Ramsey (1970) described several convolutional interleavers with arbitrary spread (s, t) and minimum latency. One of them, the *type III* (s, t) *interleaver* assures

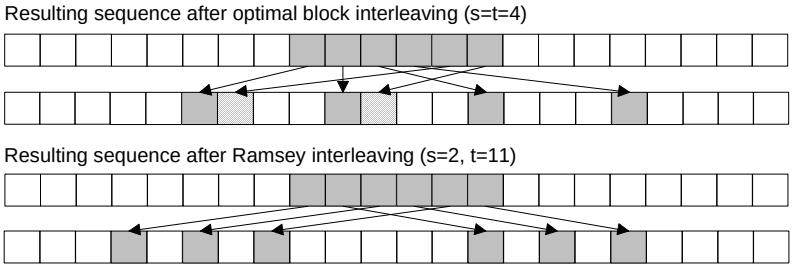


Fig. 5. Example comparing burst spreading achieved by MLBI and Ramsey-derived interleavers. Lost packets are represented in gray.

a minimum latency, given by $(s-1)(t+1)$, whenever s and t are relatively prime and $t > s$. In our previous work (Gómez et al., 2007) we showed that an invertible pair of interleavers with spread $(s = 2, t = 2B+1)$ (with $B \geq 1$) could be derived from the graphical descriptions offered by Ramsey. This invertible pair is given by the following expressions:

$$\pi(i) = i + (i \bmod 2) \cdot 2(B+1) \tag{10}$$

$$\pi^{-1}(i) = (i \operatorname{div} 2) \cdot 2 - (i \bmod 2) \cdot (2B+1) \tag{11}$$

In such a case s and t are always relatively prime and $t > s$. Therefore, the interleaver has minimal latency given by $2(B+1)$.

Since the mathematical analysis of this interleaver is detailed in (Gómez et al., 2007), here we will focus on practical considerations. Once a burst appears, the Ramsey interleaver spreads it into two *one-received-one-lost* sequences as shown in figure 5. This interleaving structure is quite robust since it grants that bursts are scattered in completely isolated losses provided their length is less or equal to t ($t = 11$ frames in the figure). In contrast, this dispersion is only granted for burst lengths of s or less frames when a block interleaver is applied (4 frames in the figure, at the same latency).

The distance between isolated losses is just one frame, but this does not involve a performance degradation when the concealment technique only requires one received frame to compute replacements for lost frames (as in the case of the mitigation algorithm proposed in the ETSI DSR standard). Instead, the main drawback of this short distance appears when two consecutive burst are considered: if these are too close an artificial burst can appear, as figure 6 shows. Fortunately, it can be proved that the artificial burst is shorter or equal to the shortest original burst implicated. We will tackle this effect in the next section.

Table 3 shows the average word accuracies (WM, MM and HM experiments) obtained by the MLBI interleaver and the proposed Ramsey-derived $(2, t)$ interleaver for the different channel conditions at several allowed delays. Plain

			MLBI Interleaving				Ramsey-derived Interleaving			
			Delay (ms)				*Delay (ms)*			
Ch.	Aur	WVA	*60*	*120*	*200*	*300*	*60*	*120*	*200*	*300*
1	85.89	85.97	86.79	86.85	86.83	86.89	86.98	86.79	87.08	87.18
2	78.95	81.47	84.06	84.72	85.54	86.05	84.95	85.44	86.01	86.09
3	71.54	75.44	79.56	80.75	82.25	83.44	81.48	82.35	83.21	83.93
4	63.48	69.43	74.38	76.67	78.02	79.45	76.85	77.95	78.59	79.64
5	54.10	60.62	65.82	68.59	70.32	73.02	68.00	70.59	72.58	73.14

Table 3
Average word accuracies obtained by WVA after MLBI and Ramsey-derived interleaving for different allowed delays in comparison with Aurora and WVA without interleaving.

WVA (nearest received repetition with no replicas) was used as EC technique whilst dynamic feature confidences were calculated by the Min-Prod strategy described in section 3. Results obtained through WVA without interleaving and Aurora standard mitigation are also included as reference. It can be observed that recognition performance can be significantly improved by means of both interleavers (also when acoustical adverse conditions are considered). However, in comparison, Ramsey-derived interleaving provides better results, particularly at low latencies, which are those we are interested in. Only when allowing a delay of 300 ms MLBI interleavers provide comparable results, but such a delay would lead to a latency of 600 ms which is possibly unsuitable for speech recognition applications.

5 VQ Replicas and Interleaving: Double Stream Scheme.

As independent techniques, there is no reason why interleaving and FEC-based schemes can not be jointly applied to achieve better results. The most immediate way to combine both techniques is to apply one technique after the other, i.e. packets can be interleaved after FEC codes have been obtained or viceversa. However, since both schemes cause a delay, this direct composition of FEC and interleaving processes would result in a sum of their corresponding delay (and latency).

This sum of delays can be prevented by a double stream (DS) scheme (Gomez et al., 2007). In this scheme, two independent sequences of feature vectors are considered: a primary stream which consists of feature vectors that are quantized by SVQ as in the DSR standard, and a secondary stream composed of the same vectors but VQ quantized. This second flow is just a redundant stream inserted within the packets by means of VQ replicas. Different interleaving functions are applied to different streams so that the total latency

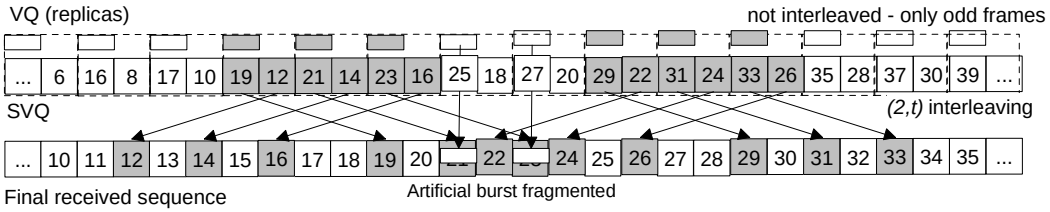


Fig. 6. Example of the proposed DS scheme where an artificial burst (from vector 21 to 24) caused by the interleaving over two different bursts (gray frames) is re-fragmented thanks to the replicas. Each packet is composed of two $(2, t)$ interleaved SVQ vectors and one VQ replica (only odd vector, no interleaving).

caused in the transmission is equal to the maximum of both interleavers.

This DS scheme opens a number of possibilities since now two interleavers must be chosen in order to jointly maximize the spreading of losses considering both SVQ and VQ vectors. A relevant hurdle to be considered is the appearance of overlapped replicas. There exist combinations of interleavers which cause that some replicas are sent in the same packet that the vector to be protected. Those replicas are useless, since both the vector and its replica will be lost if the packet is not received.

Due to its complexity, we tackle this problem in two steps: first by selecting a suitable SVQ interleaver and then a VQ one which, at least, avoids the presence of overlapped replicas. Since it has provided the best results, the $(2, t)$ interleaver described in section 4 and defined by equations (10) and (11), is applied as primary interleaver (i.e. to reorder the SVQ vectors). This interleaver will define the total latency of the scheme, so its parameter t is set according to the maximum allowed delay. For the secondary interleaving, previously described re-orderings could be considered but, unfortunately, both the MLBI interleavers and the ordering initially proposed for VQ replicas (section 3.3) lead to overlapped replicas. A detailed analysis of the first ones reveals the appearance of overlapped replicas at some time instants, while the second ones can be considered as a particular case of the DS scheme where SVQ vectors are not interleaved and VQ replicas are reordered using a $(2, t)$ interleaver with $t = T_{fec}$ and then inserted within the packets. Thus, this reordering would also lead to overlapped replicas.

Curiously, not to interleave any replica at all can also be a good solution. Interleaving, as well as it can spread bursts, can also build up artificial bursts from near losses or bursts (due to the reordering). In the case of the Ramsey-derived interleaver, it is possible that two close bursts can be spread in such a way that an artificial consecutive loss appears as a consequence of these (figure 6). By not interleaving the replicas or by interleaving them with a short spread, these artificial bursts can be fragmented.

In our proposal, we take advantage of this along with the redundant nature

			DS scheme (wo. discard) 2 x 4bit replica, delay (ms)				DS scheme (w. discard) 1 x 8bit replica, delay (ms)			
Ch.	Aur	WVA	*60*	*120*	*200*	*300*	*60*	*120*	*200*	*300*
1	85.89	85.97	87.00	86.84	87.12	87.20	87.08	87.09	87.18	87.24
2	78.95	81.47	85.14	85.74	86.17	86.19	85.37	85.82	86.45	86.42
3	71.54	75.44	81.82	83.09	83.79	84.47	82.16	83.53	84.37	84.84
4	63.48	69.43	77.12	79.18	80.14	80.86	77.98	79.66	80.85	81.90
5	54.10	60.62	68.69	72.48	74.64	74.21	70.21	73.16	75.32	76.18

Table 4
Average word accuracies obtained by the Double Stream scheme (with and without discarded replicas) for different allowed delays in comparison with Aurora and the plain WVA technique.

of replicas. This last feature allows that the interleaving function applied over replicas does not need to be a permutation any longer, so that some replicas could be discarded allowing others to be better represented. Thus, we send the VQ replicas in their original order (no interleaving) but only half of them (one replica per packet). That is, all redundancy bits available (n) are devoted to odd replicas while even replicas are not transmitted (viceversa is also valid). Given the high time correlation between vectors, the results are almost the same as transmitting all the replicas at a double bit-rate ($n + n$).

This scheme, depicted in figure 6, has shown the best results in comparison with other lossless interleavers tested, i.e. two replicas per packet with ($n/2 + n/2$) bits. Table 4 shows the average results, over the WM, MM and HM experiments, achieved by our proposed scheme with and without discarding replicas under different channel conditions. As can be observed, the additional information contained in the secondary stream (or VQ replicas) allows an improvement on word accuracy in comparison with WVA schemes with interleaving (table 3). It can be noted that this performance increase (from WVA to WVA with VQ) is not so high as when there was no interleaving (table 2). The reason is that bursts have been already broken by Ramsey-derived interleaving, so that the only goal of replicas is now to recover lost information and re-fragment artificial bursts. Figure 7 details the performance of our proposal (WVA plus interleaving and only one VQ replica per packet) in the WM, MM and HM experiments in comparison with the previous techniques described in this paper. For all of them, a maximum delay of 120 ms has been allowed, as well as a possible bandwidth overhead of 8 bits (that can be introduced within the payload at no effective cost). As it is observed, our scheme significantly increases the robustness against packet losses both in mismatch and well-matched conditions.

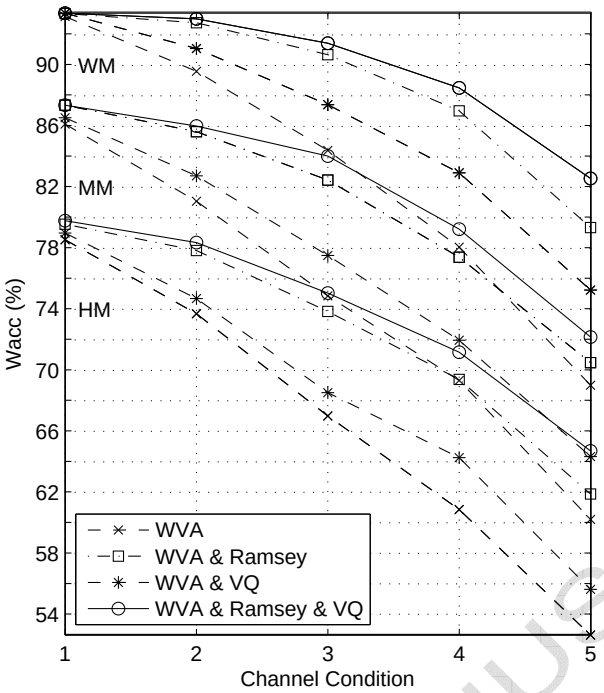


Fig. 7. Comparison between WVA with no sender-driven technique (WVA), WVA with VQ replicas (WVA & VQ) or Ramsey-derived interleaving (WVA & Ramsey) and the proposed DS scheme (WVA & Ramsey & VQ) for WM, MM and HM experiments under different channel conditions. Same allowed delay (120 ms) and bandwidth overhead (8 bits) have been considered for all techniques.

6 Conclusions

In this paper, we propose a whole recovery scheme designed to improve the robustness against packet losses on distributed speech recognition systems. In this scheme, two sender-driven techniques, namely, media-specific FEC codes and frame interleaving, are jointly applied along with a receiver-based EC technique, the Weighted Viterbi algorithm. The sum of the delays caused by the use of distant replicas and interleaving is avoided by means of a double stream scheme. In this scheme, two different flows, one consisting of SVQ vectors and the other composed of VQ vectors or replicas are considered. Before replicas are inserted within the packets as FEC codes, both streams are independently interleaved, so that the latency is limited by the maximum delay of both interleavers.

As primary interleaver, we propose a $(2, t)$ convolutional interleaver. We have shown that, at the same latency, this Ramsey-derived interleaver can disperse into isolated losses longer bursts than the classical MLBI interleaving. This is possible since the distance between the scattered losses is just one frame ($t = 2$ instead of $t = s$), which is long enough for the usual reconstruction techniques applied in DSR and, in particular, for those considered in this paper (i.e. near-

est vector repetition). As secondary interleaver, we apply a lossy reordering function. That is, taking advantage of the redundant nature of replicas and the high correlation between consecutive frames, not all VQ vectors are transmitted. In each packet, all available bits are devoted to representing only the odd frame (or, alternatively, the even one) which is sent in its corresponding time-instant (no reordering). In addition to the introduction of additional information, these replicas allow to re-fragment the possible artificial bursts caused by the primary interleaver.

At the receiver, received and lost vectors and replicas are differently treated by the recognizer by means of the WVA algorithm. A reliability function based on the normalized cross covariance is used as static feature confidence estimate. This reliability function can offer a confidence estimate for a recovered lost feature given the reconstruction method, including those based on VQ replicas. Thus, the recognizer can be informed that we have some information about lost frames, the VQ replicas, but that our confidence in them is lower, as they have been degraded by a strong quantization process. Confidence values for dynamic features are calculated from the reliabilities of the static features within the sliding window from which those are calculated. Although frequently forgotten in the WVA algorithm, delta features provide a speech representation more robust against acoustic noise. Confidences for first derivatives are obtained as the minimum of the static ones, while for second derivatives are obtained as their product. It has been shown that this approach, which mixes minimum and product strategies, achieves good results in both well-matched and mismatched acoustic conditions.

The proposed scheme and techniques have been tested under different channel conditions and considering several levels of acoustic mismatch between training and testing sets (high, medium and no mismatch). In all cases our proposal provides a considerable improvement on recognition accuracy. Thus, assuming a reasonable delay of 120 ms and reusing the zero padding and CRC bits (8 bits in total), up to a 44.3% of relative reduction on the average word error rate can be achieved in comparison with the Aurora standard, and 25.7% and 33.5% in comparison with a WVA system with and without replicas, respectively. In general, the improvements achieved have shown a high degree of independence from the acoustic environment. In this sense, the techniques described in this paper provide comparable performance improvements under high and medium acoustic mismatch and well-matched experiments.

References

Andrews, K., Heegard, C., Kozen, D., 1997. A theory of interleavers. Technical report.

Bernard, A., Alwan, A., 2002. Low-bitrate distributed speech recognition for

packet-based and wireless communication. IEEE Transactions on Speech and Audio Processing 10, 570–579.

Bolot, J., 1993. End-to-end packet delay and loss behavior in the Internet. ACM Sigcomm, 289–298.

Cardenal-Lopez, A., Docio-Fernandez, L., Garcia-Mateo, C., 2004. Soft decoding strategies for distributed speech recognition over IP networks. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP), 49–52.

Cardenal-Lopez, A., Garcia-Mateo, C., Docio-Fernandez, L., 2006. Weighted viterbi decoding strategies for distributed speech recognition over IP networks. Speech Communication 48, 1422–1434.

Endo, T., Kuroiwa, S., Nakamura, S., 2003. Missing feature theory applied to robust speech recognition on IP networks. Proceedings of Eurospeech.

ETSI-ES201-108, 2000. Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.

ETSI-ES202-050, 2002. Advanced front-end feature extraction algorithm.

ETSI-ES202-211, 2003. Distributed Speech Recognition; Extended Front-end Feature Extraction Algorithm; Compression Algorithm, Back-end Speech Reconstruction Algorithm.

ETSI-ES202-212, 2005. Distributed Speech Recognition; Extended Advanced Front-end Feature Extraction Algorithm; Compression Algorithm.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. on ASSP 34, 52–59.

Gómez, A., Peinado, A., Sánchez, V., Milner, B., Rubio, A., 2004. Statistical-based reconstruction methods for speech recognition in IP networks. COST 278 - Robust2004: Robustness Issues in Conversational Interaction.

Gómez, A., Peinado, A., Sánchez, V., Rubio, A., 2003. A source model mitigation technique for distributed speech recognition over lossy packet channels. Proceedings of EuroSpeech.

Gómez, A., Peinado, A., Sánchez, V., Rubio, A., 2006. Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels. IEEE Trans. on Multimedia, 1228–1238.

Gomez, A., Peinado, A., Sanchez, V., Rubio, A., 2007. An integrated scheme for robust distributed speech recognition over lossy packet networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4, IV–857–IV–860.

Gómez, A., Peinado, A., Sánchez, V., Rubio, A., 2007. On the Ramsey class of interleavers for robust speech recognition in burst-like packet loss. IEEE Trans. on Speech and Audio Processing, 1496–1499.

Ion, V., Haeb-Umbach, R., 2006. Uncertainty decoding for distributed speech recognition over error-prone networks. Speech Communication 48, 1435–1446.

James, A., Gómez, A., Milner, B., 2004. A comparison of packet loss compensation methods and interleaving for speech recognition in burst-like packet

loss. Proc. International Conference on Spoken Language Processing (ICSLP).

James, A., Milner, B., 2004. An analysis of interleavers for robust speech recognition in burst-like packet loss. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP).

James, A., Milner, B., 2005. Soft decoding of temporal derivatives for robust distributed speech recognition in packet loss. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP), 345–348.

Koodli, R., Ravikanth, R., 1999. One-way loss pattern sample metrics. RFC 3357.

Macho, D., 2000. Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End evaluation: Description and baseline results. URL `http://www.elda.fr`

Milner, B., James, A., 2003. Analysis and compensation of packet loss in distributed speech recognition using interleaving. Proceedings of Eurospeech.

Milner, B., James, A., 2004. Packet loss modelling for Distributed Speech Recognition. COST 278, Robust2004: Robustness Issues in Conversational Interaction.

Milner, B., James, A., 2006. Robust speech recognition over mobile and IP networks in burst-like packet loss. IEEE Transactions on Audio, Speech and Language Processing 14, 223–231.

Milner, B., Semnani, S., 2000. Robust speech recognition over IP networks. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP) 3, 1791–1794.

Peinado, A., Gómez, A., Sánchez, V., Perez-Cordoba, J., Rubio, A., 2006. An integrated solution for error concealment in DSR systems over wireless channels. Proceedings of INTERSPEECH.

Peinado, A., Gómez, A., Sánchez, V., Rubio, A., 2005. Packet loss concealment based on VQ replicas and MMSE estimation applied to Distributed Speech Recognition. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP), 329–332.

Peinado, A., Segura, J., 2006. Speech Recognition over Digital Channels, Robustness and Standards. Wiley and Sons Ltd.

Perkins, C., Hodson, O., Hardman, V., 1998. A survey of packet-loss recovery techniques for streaming audio. IEEE Network Magazine.

Postel, J., 1980. User datagram protocol. RFC 768.

Potamianos, A., Weerackody, V., 2001. Soft-feature decoding for speech recognition for wireless channels. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP).

Ramsey, J., 1970. Realization of optimum interleavers. IEEE Trans. on Information Theory 6, 338–45.

Schulzrinne, H., Frederick, R., Jacobson, V., 1996. RTP: A Transport Protocol for Real-Time Applications. RFC 1889.

Xie, Q., Pearce, D., Balasuriya, S., Kim, Y., Maes, S., Garudari, H., 2002. RTP payload format for DSR ES 201 108. IETF Audio Video Transport

WG, Internet RFC 3557.

Yoma, N., McInnes, F., Jack, M., 1998. Weighted viterbi algorithm and state duration modeling for speech recognition in noise. Proc. IEEE Int.Conf. on Acoustics, Speech and Signal Processing (ICASSP).