interspeech 2008 incorporating SST 08

Intelligibility Evaluation of Ramsey-Derived Interleavers for Internet Voice Streaming with the iLBC Codec

Angel M. Gomez, José L. Carmona, Antonio M. Peinado, Victoria Sánchez, José A. Gonzalez

Dpt. Teoría de la Señal, Telemática y Comunicaciones,
University of Granada, Spain
{amgg,maqueda,amp,victoria,joseangl}@ugr.es

Abstract

This paper focuses on the application of a previously proposed interleaving, derived from the Ramsey convolutional class, in a voice streaming context with bursty packet losses. This kind of interleaving has already shown significant improvements in a distributed speech recognition context, in comparison with the widely used minimum latency block interleavers (MLBI). Here, the effectiveness of these interleavers is evaluated with an internet-oriented speech codec, such as iLBC. Since iLBC avoids error propagation due to lost frames and only uses the previously received frame to recover from those, Ramsey interleaving turns out especially suitable for this codec. In order to measure the performance of the system, the ITU PESQ algorithm is applied along with an intelligibility criterion based on the accuracy obtained through an automatic speech recognizer (ASR). In addition, an informal subjective test is carried out to corroborate the ASR scores. Results show that the proposed Ramsey-derived interleaving provides at least the same quality and better intelligibility than the MLBI ones when it is applied to iLBC codec.

1. Introduction

Bit errors can be neglected in IP networks for a number of reasons. Reliable underlying networks, error checking in UDP headers and detection and/or correction mechanisms usually included within the payload could make them seem as noiseless transmission media. However degradation appears due to the drawbacks inherent in its packet-switching structure, which are mainly late packets, time spreading and packet loss. Since the first two can be easily treated by means of the introduction of decoding delays, packet losses appear as the main source of degradation.

Unfortunately, lost packets tend to appear consecutively (i.e. in bursts), causing a more negative impact. It is well established that in speech related applications (such as speech transmission, voice streaming, distributed speech recognition, etc.) packet losses are more harmful when they are consecutive [1, 2, 3]. The reason is that error correction and concealment techniques can be quite effective when the consecutive packet losses are short (i.e. more random) but are not so effective for long bursts. Thus, the subjective quality degradation increases as the burst length increases.

Based on this fact, robustness against bursts of losses can be increased by applying an *interleaver* prior to transmission. Interleaving does not modify the packet loss ratio. Instead, losses are shaped in a less damaging distribution. By means of a reordering of the speech frames, interleaving reduces the burst length at the receiver, allowing the concealment technique to perform better and improving the subjective quality. As an advantage, interleaving does not increase the required bandwidth, although it causes a latency in the transmission.

Minimum latency block interleavers (MLBI) [4] are a widely used class of interleavers. However, in this work we describe a different kind of interleaving derived from the Ramsey convolutional class [5]. In a previous work [3], we showed that significant improvements could be achieved with Ramsey-derived interleavers (in comparison with MLBI ones) when they were applied to a distributed speech recognition system. The main idea is that their parameters are more flexible and could be better adjusted to the particularities of the concealment technique. Here, we test their performance in a voice transmission and/or streaming context over IP networks. In order to do so, the recently proposed internet-oriented iLBC codec [6] is used. This codec avoids inter-frame dependencies so that errors caused by frame losses are not propagated. This feature makes Ramsey-derived interleaving specially suitable for iLBC codec, as we will show.

This paper is organized as follows: first, some concepts of frame interleaving, and MLBI and Ramsey-derived interleavers, are briefly explained. Then, in section 3, iLBC principles are described and the suitability of Ramsey interleaving for this codec justified. Section 4 is devoted to the experimental framework, while results are shown in section 5. Finally, conclusions are summarized in section 6.

2. Frame-Level Interleaving

Interleaving is a technique commonly applied at the bit level to randomize the appearance of errors, thus reducing the effect of error bursts. At this level, interleaving is useless to IP networks, since errors comprise complete packets of information (which are lost). However, interleaving can also be applied to larger transmission units. In particular, a useful approach is to interleave speech frames [2, 7]. We will refer to these interleavers as frame-level interleavers.

Frame-level interleavers permute the order in which complete frames are transmitted, so that, when the original order is restored, consecutive losses appear scattered at the receiver. The ability of an interleaver to disperse consecutive losses (or errors) is related to its *spread*. An interleaver π has spread (s, t) if any two input symbols in an interval of length s are scattered by a distance of at least t symbols at the output [4]. It can be shown that, if a burst of frame losses with length less than t appears, an interleaver with spread (s, t) will disperse it into

This work was funded by the project MEC/FEDER TEC2007-6660.

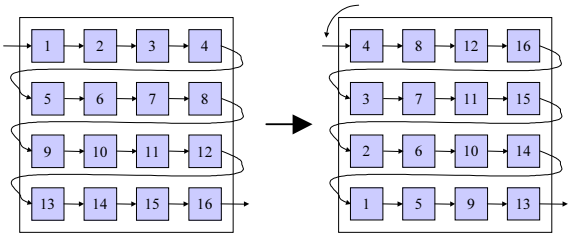

Figure 1: *Illustration of a $s = 4$ block interleaver (equivalent to a rotation of 90º anti-clockwise).*

isolated frame losses separated by at least $s - 1$ frames [3].

As can be observed, large values for s and t are desirable. However, these ones entail longer latencies. The latency of an interleaver (l_π) is given by the sum of two delays: a) the maximum time delay that an input symbol waits in the interleaver until it is produced as an output, plus b) the minimum delay required by the interleaver to become realizable (details can be found in [4] and [3]). As expected, interest is focused on finding those interleavers which provide the maximum spread causing the shortest possible latency.

2.1. Minimum latency block interleavers

An extensively applied class of interleavers are block interleavers. A block interleaver of period p operates in blocks of p elements, permuting these elements among themselves. Most of the popularity of these interleavers relies on their easy implementation and the possibility of achieving minimum latency. Thus, it can be proved that there are two block interleavers which have minimal latency among all block interleavers of spread (s, s) (or simply s) [4]. These are given by,

$$\pi_1(is + j) = (s - 1 - j)s + i \qquad 0 \le i, j \le s - 1, \qquad (1)$$

$$\pi_2(is + j) = js + (s - 1 - i) \qquad 0 \le i, j \le s - 1. \qquad (2)$$

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 speech frames either 90º clockwise or 90º anticlockwise (as shown in figure 1). Thus, in practice, they can be implemented through a matrix where the input data is written along the rows and read out along the columns. The latency introduced by minimum latency block interleavers (MLBI) is related to their spread and is equal to $l_\pi = 2s(s - 1)$ frames.

2.2. Ramsey's convolutional interleavers

In contrast to the block ones, convolutional interleavers do not satisfy that if p is a period of π, there is an interval of length p whose image under π is also an interval of length p. In practice, this implies that convolutional interleavers are not shift equivalent (i.e. equivalent with a possible delay) to a permutation interleaver [4], making their mathematical analysis more difficult.

In [5] Ramsey provides a method to build up convolutional interleavers with spread (s, t) and minimum latency. This minimum latency is given by $(s-1)(t+1)$ and it is achieved when some primeness conditions are satisfied [5]. The main problem with MLBI interleaving is that t is assumed to be equal to s (spread (s, s)). Depending on the mitigation technique, such a distance between isolated losses can be useless, increasing the latency of the interleaver without any real advantage. As we showed in a previous work [3], in the case of distributed speech

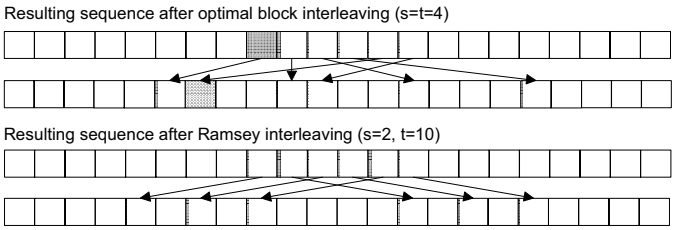

Figure 2: *Comparison example of burst spreading using MLBI and Ramsey-derived interleavers. Lost packets are represented in gray.*

recognition it was significantly better to set s as short as 2, increasing t along with the allowed delay. As a result, we proposed the following invertible pair of interleavers, derived from the Ramsey's type III (s, t) interleaver:

$$\pi(i) \quad = i + (i \bmod 2) \cdot 2(B + 1) \qquad (3)$$

$$\pi^{-1}(i) \quad = (i \operatorname{div} 2) \cdot 2 - (i \bmod 2) \cdot (2B + 1) \qquad (4)$$

where $s = 2$ and $t = 2B+1$ ($B \ge 1$). Since s and t are relative primes and $t > s$, the interleaver has minimal latency which is given by $l_\pi = 2(B + 1)$.

The mathematical analysis and justification of this interleaver is detailed in [3]. Here we will focus on practical considerations. Given a burst, this interleaver spreads it into two *one-received-one-lost* sequences as shown in figure 2. This structure is quite robust 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 DSR standard from ETSI [8]). As can be seen in figure 2, Ramsey-derived interleaving grants that bursts are scattered in completely isolated losses provided their length is less or equal to t (10 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.

3. Suitability of Ramsey interleavers for iLBC Codec

The Internet Low Bitrate codec (iLBC) is a recently proposed speech codec which, in contrast to others commonly applied to packet-switched networks, avoids any inter-frame dependence. It is well known that, when long-term (LTP) and short-term predictive filters (LP) are used, de-synchronization between memories of filters appears after a missing frame, and errors can propagate during several correctly received frames [9, 10]. In the iLBC codec, each frame is encoded separately from adjacent frames, so that, although predictive procedures are used, they do not extend beyond frame boundaries. Thus, the error caused at the decoder by a lost frame is not propagated [9] achieving an increase of the perceptual quality. As a disadvantage, the codec has higher bit-rates than other speech codecs: 15.2 kbps for the 20 ms mode and 13.3 kbps for the 30 ms (in this paper we focus on the first one).

On the other hand, the packet loss concealment (PLC) algorithm proposed for iLBC is based on a pitch-synchronous repetition of the excitation signal, which is filtered by the last LP filter of the previous block. The excitation in the previous block is used to create the excitation for the block to be substituted, while a correlation analysis also performed on the previous block detects the pitch value. This procedure is repeated in

a similar manner but decreasing the energy of the excitation, i.e. muting the output signal progressively.

From these features, it can be derived that forcing the interleaver to grant consecutive received frames after a loss (at the cost of counteract shorter bursts) is almost useless with this codec. On one hand, only the previous block is needed to provide a replacement for lost frames. On the other, since coding independence among frames is achieved in iLBC, a consecutive reception of frames is not needed to recover from filter desynchronization as in other codecs. These facts make the *one-received-one-lost* sequence, provided by a Ramsey interleaver (with $s = 2$), a preferable frame loss distribution for this codec, especially when it is considered that such a distribution allows to counteract longer bursts (at the same latency).

4. Experimental Framework

In this paper we evaluate the performance, mainly in terms of intelligibility, of the Ramsey-derived interleavers applied to the iLBC codec. In order to do so, the performance obtained through an automatic speech recognizer (ASR) is used as objective intelligibility measure. It has been observed that, in noise-free conditions, the automatic speech recognition accuracy is highly correlated to human intelligibility [11]. Thus, an idea of the speech intelligibility can be obtained through the accuracy given by an ASR system. However, since it is clear that this method is rather indirect and potentially prone to some side effects, an informal subjective evaluation of intelligibility has been also performed to check the results.

In addition, an objective method to evaluate the speech quality has been applied. Since the signal before transmission is available, the ITU recommendation P.862 [12], also known as PESQ (Perceptual Evaluation of Speech Quality) algorithm, has been used.

4.1. Channel Simulation

The channel burstiness exhibited by IP networks is modeled by a 2-state Markov model [13]. The model parameters can be set in accordance with an average burst length (L_{loss}) and a loss ratio (R_{loss}). Five channel conditions are proposed with loss packet ratios of 10%, 20%, 30%, 40% and 50% with an average burst length of 1, 2, 4, 6 and 8 packets, respectively. Although later conditions may show unrealistically high amounts of packet loss, the only purpose of this is to provide a significant number of bursts, since fewer and fewer bursts appear as L_{loss} increases.

4.2. Objective intelligibility evaluation

For the intelligibility evaluation we use the Aurora-2 database [8]. This database consists of utterances with connected digits. The vocabulary is made up of 11 digits between 0 and 9 (zero has two sound descriptions: 'zero' and 'o'). A feature extractor segments the decoded speech signal into overlapped frames of 25 ms every 10 ms. Each speech frame is represented by a feature vector containing 13 Mel Frequency Cepstrum Coefficients plus the log-Energy and then extended with their first and second derivatives.

The speech recognizer is based on Hidden Markov Models (HMM). It uses eleven 16-states continuous HMM word models (plus silence and pause, which have 3 and 1 states, respectively) with 3 gaussians per state (except silence which has 6 gaussians per state). The training and testing data are extracted from the Aurora-2 database. The training is performed with 8440 clean sentences while tests are carried out over the 4004 clean sentences of *set A*. Word accuracy ($W_{acc}(\%)$) is chosen as intelligibility measure.

4.3. Subjective intelligibility evaluation

Informal subjective evaluation involved 15 listeners who evaluated at least two utterances per channel condition. Since all of them are Spanish native speakers, utterances were selected from the Spanish subset of Aurora-3 database. As Aurora-2, this database consists of utterances with connected digits which have a vocabulary of 10 digits, from 0 to 9. In order to limit the number of tests, MLBI and Ramsey-derived interleavers were tested at the only latency of 6 speech frames (120 ms). Packet losses were randomly generated, according to the channel model described in subsection 4.1, for each listener, utterance and channel condition. However, in order to reduce undesired variability, in each experiment both interleavers scattered exactly the same random loss pattern. As before, the mean of the word accuracy ($W_{acc}(\%)$) achieved by the listeners is chosen as intelligibility measure.

4.4. Perceptual quality evaluation

A Perceptual quality evaluation score is obtained through the PESQ algorithm. As in the recognition tests, set A from the Aurora-2 database is used. However, original test utterances were concatenated into groups of seven, resulting in a total of 572 sentences. The reason for this is that PESQ algorithm has not been designed to evaluate short sentences [12]. Lengths between 8 and 20 s are recommended, but Aurora-2 utterances have a mean duration of only 1.5 s. Through this grouping, mean duration is extended to 12 s (approx.), with minimum and maximum values of 7.5 s and 20 s respectively. In order to obtain an overall score for the tested condition, the score of each sentence is weighted by its length.

5. Experimental results

Table 1 shows the results obtained with the iLBC codec applying Ramsey-derived and MBLI interleaving under the proposed channel conditions. Latencies allowed to the interleavers are 6, 12 and 20 speech frames (120, 240 and 400 ms, respectively). Two scores are shown, the word accuracy offered by the ASR recognizer (cols. 2 and 9) and the one provided by the PESQ algorithm (cols. 9-15). In addition, results obtained without interleaving are also included as baseline. Recognition word accuracy under a clean channel condition is 99.02 %, while PESQ score is 3.94 for iLBC.

In terms of quality, both interleavers achieve comparable PESQ scores. Only slight differences are observable between them, being the largest difference lower than 0.1. Thus, although PESQ scores and subjective quality have a limited correlation, it seems clear that the Ramsey interleaver at least provide the same perceptual speech quality as MLBI ones.

In contrast, in terms of intelligibility, Ramsey-derived interleavers clearly provide better speech recognition scores (word accuracy) than MLBI ones, particularly at low latencies. Assuming a latency of 6 frames, the differences between both interleavers are notable. These differences reduce as allowed latency increases. This can be explained by the fact that, although Ramsey interleavers can counteract longer bursts than MLBI, longer and longer bursts become scarce in the tested conditions (maximum L_{loss} is 8 packets). Thus, as allowed latency increases, the performance of both interleavers becomes simi-

Ch.	*Speech Recognition Word Accuracy (Wacc %)*							*ITU PESQ score (-0.5 – 4.5)*						
	Base	MLBI			Ramsey-derived			Base	MLBI			Ramsey-derived		
		6	12	20	6	12	20		6	12	20	6	12	20
1	96.22	97.58	97.66	97.68	97.71	97.76	97.64	3.15	3.28	3.28	3.27	3.26	3.26	3.26
2	86.93	93.47	94.70	95.19	95.11	95.27	95.18	2.66	2.87	2.90	2.90	2.85	2.84	2.84
3	72.52	82.43	86.59	88.99	87.29	89.20	89.08	2.26	2.47	2.53	2.56	2.53	2.52	2.53
4	59.32	69.83	75.54	79.71	77.49	79.98	80.14	1.95	2.13	2.20	2.24	2.21	2.22	2.23
5	48.10	58.98	65.51	71.19	66.77	71.12	71.44	1.72	1.87	1.96	1.99	1.96	1.99	1.99

Table 1: *Results obtained by the iLBC codec with Ramsey-derived and MLBI interleaving (latencies of 6, 12, 20 frames) under the proposed channel conditions.*

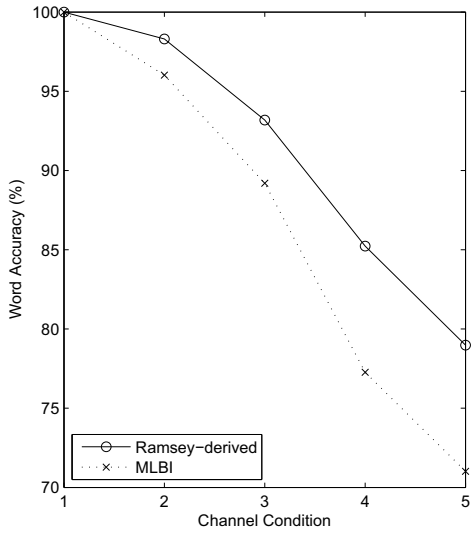

Figure 3: *Mean word accuracy obtained by human listeners for Ramsey-derived and MLBI interleavers (latency of 6 frames) under the proposed channel conditions.*

lar. However, this does not happen until latencies as long as 20 frames (400 ms) which normally are undesirable.

The indirect results on intelligibility provided by the speech recognizer are confirmed by results in figure 3. This figure shows the mean word accuracy achieved by human listeners and gives us an actual idea of the performance of both interleavers in terms of intelligibility. As can be observed, significantly more words can be understood by a human listener when a Ramsey-derived interleaver with a latency of 6 frames is used in comparison with an MBLI one (at the same latency).

6. Conclusions

In this paper we describe and evaluate a pair of frame interleavers derived from the Ramsey convolutional class. In contrast to the widely used minimum latency block interleavers, Ramsey interleavers allow to independently control the spread parameters s and t, both related to the latency of the interleaver. By fixing s as low as $s = 2$ (and increasing t as much as latency allows), a spreading of the form one-received-one-lost is achieved. This burst dispersion is particularly useful in distributed speech recognition systems based on the ETSI standard, as we showed in a previous work, but also in a voice streaming context over a packet-switching network using the internet-oriented iLBC codec.

By means of three different tests, we have shown that, in comparison with MLBI, Ramsey-derived interleavers allow iLBC to achieve better results in terms of intelligibility while perceptual quality is maintained. Quality evaluation has been performed by means of the ITU PESQ algorithm while the performance obtained through an automatic speech recognizer has been used as intelligibility criterion. Since this method can be considered prone to side effects, an additional subjective intelligibility evaluation with human listeners has also been performed, confirming the results obtained by ASR.

7. References

[1] Y.J. Liang and J.G.B. Apostolopoulos, "Model-based delay-distortion optimization for video streaming using packet interleaving," *IEEE Asilomar Conference on Signals, Systems and Computers*, pp. 1315–1319, 2002.

[2] C.Perkins, O.Hodson, and V.Hardman, "A survey of packet-loss recovery techniques for streaming audio," *IEEE Network Magazine*, 1998.

[3] A.M. Gómez, A.M. Peinado, V. Sánchez, and A.J. Rubio, "On the ramsey class of interleavers for robust speech recognition in burst-like packet loss," *IEEE Trans. on Audio Speech and Language Processing*, vol. 8, pp. 1496–1499, 2007.

[4] K. Andrews, C. Heegard, and D. Kozen, "A theory of interleavers," *Technical report 97-1634*, 1997.

[5] J. Ramsey, "Realization of optimum interleavers," *IEEE Trans. on Information Theory*, vol. 6, pp. 338–45, 1970.

[6] S.V. Andersen, W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, and J. Skoglund, "iLBC - a linear predictive coder with robustness to packet losses," *IEEE Workshop Proceedings. on Speech Coding*, pp. 23–25, October 2002.

[7] 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.

[8] 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.

[9] R. Lefebvre, P. Gournay, and R. Salami, "A study of design compromises for speech coders in packet networks," *IEEE International Conference on Speech and Signal Processing*, vol. I, pp. 265–269, 2004.

[10] A.M. Gómez, A.M. Peinado, V. Sánchez, and A.J. Rubio, "Recognition of coded speech transmitted over wireless channels," *IEEE Trans. on Wireless Comunications*, 2006.

[11] W. Jiang and H. Schulzrinne, "Speech recognition performance as an effective perceived quality predictor," *Tenth IEEE International Workshop on Quality of Service*, pp. 269–275, May 2002.

[12] *Perceptual evaluation of speech quality (pesq): An objetive method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs*, ITU-T P.862 Recommendation, 2001.

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