Intelligibility Evaluation of a VoIP Multi-flow Block Interleaver*

Juan J. Ramos-Muñoz, Ángel M. Gómez, and Juan M. Lopez-Soler

Signal Theory, Telematics and Communications Department
University of Granada (SPAIN)
jjramos@ugr.es, amgg@ugr.es, juanma@ugr.es

Abstract. This work contributes to demonstrate what perceptual benefits can be expected by adding some processing capabilities to the network nodes for the class of interactive audio streaming applications. In particular, we propose a new voice stream multi-flow block interleaver and we show that it provides an intelligibility performance very close to the reference end-to-end interleaver, even under conditions where the end-to-end interleaving is unfeasible.

Keywords: VoIP subjective evaluation, interleaving, speech recognition

1 Introduction

It is well established that in streaming voice applications packet losses are more harmful as they are consecutive, since the subjective quality degradation increases as the burst length increases [1]. Based on this fact, to improve the VoIP perceived quality some procedures should be considered to combat the unwished bursty-error-prone nature of the Internet. To this end, by using active technology, we aim to scatter the pattern of losses without increasing both the bandwidth consumption and, ideally, the end-to-end delay as well.

Traditionally, error control techniques operate end-to-end [2]. However, with the advent of the Active Networks technology [3] new promising router functionalities can be envisaged.

In this work, we evaluate our active procedure in terms of intelligibility, and we experimentally show that the quality of the received audio stream increases by using active routers. More precisely, we take up again the packet interleaving problem but now we use the processing capabilities of the network elements. Since intermediate network nodes can use different multimedia flows [4], we propose an interleaver algorithm that take advantage of it [5]. For comparison purposes, we simulate a reference single flow end-to-end interleaver and, we experimentally demonstrate that, under some circumstances, our algorithm outperforms the reference system in terms of intelligibility with light impact in the end-to-end packet delay.

* This work has been partially financed by the Spanish Science and Technology Ministry under Research Project TIC2002-02978 (with 70% of FEDER funds).

2 Basic and Multi-flow Block Interleavers Algorithms

Given a input packet sequence, denoted by $\{a_i\}$, and the output sequence, denoted by $\{b_i\}$ the interleaver defines a permutation $\pi : \mathbb{Z} \mapsto \mathbb{Z}$ such that $a_i = b_{\pi(i)}$. An interleaver is said to be *periodic* if it verifies that $\pi(i+e) = \pi(i)+e$, being e its period. An interleaver has a *spread* s, if any two input symbols in an interval of length s are separated by distance of at least s at the output.

In interactive VoIP applications, the end-to-end delay must be bounded. Thus, for getting the minimum delay packet reallocation, given a spread s, the basic block interleaver algorithm [6] (hereafter referred to as *Type I* (s)) is:

1. Arrange the symbols associated to the input packets in a $(s \times s)$ matrix in rows, from left to right and from top to bottom.
2. Read the matrix by columns from bottom to top and from left to right, and accordingly send the packets.

In this case, D_{max}, defined as the maximum number of symbols (in the worst case) that any packet will wait at the interleaver, will be equal to $D_{max} = s \cdot (s-1)$. Note that for a periodic VoIP flow, given the imposed end-to-end delay constraint, *Type I* (s) interleaver will be limited to s such as $s \cdot (s-1) \cdot t_f < d_{max}$, where t_f is the inter-packet period, and d_{max} is the maximum end-to-end delay that any packet can tolerate. For example, for typical values of $d_{max} = 300$ ms and $t_f = 20$ ms, it can not be assured that bursts length longer than 4 packets will be scattered. However, if the router jointly interleaves $n_f > 1$ different flows, a reduction in the interleaver packet delay can be potentially achieved.

In this work, we will assume that all the flows have the same period t_f. To fill the interleaver matrices, each flow will maintain the relative order with respect to the others. In addition, for a given flow each row will be written from left to right according to the packet sequence number.

Let $(f^1, f^2, \ldots, f^{n_f})$ be the n_f available audio flows, and let s be the maximum expected burst length. To simplify, let us additionally define R_j^i, with $i = \{1, \ldots, n_f\}$ and $\{j = 1, \ldots, n_m\}$, as the number of consecutive rows that the flow f^i will be assigned for filling the interleaver matrix j, being n_m the number of matrices. Depending on n_f and s, we will consider two different cases.

1. Whenever $n_f \geq s$, the interleaver will be based on just one $(n_f \times 1)$ matrix $(n_m = 1)$, in which $R_1^i = 1, \forall i = \{1, \ldots, n_f\}$. For this case, the interleaver output will be given by ..., f_i^1, f_j^2, ..., $f_k^{n_f}$, f_{i+1}^1, f_{j+1}^2, ..., $f_{k+1}^{n_f}$, ..., where the subscripts i, j, ..., k denote the sequence number for flows f^1, f^2, ..., f^{n_f}. We refer to this interleaver as *Type II* (n_f).
2. If $n_f < s$, we will refer to this interleaver as *Type II* (n_f, s). Under this condition, two different cases will be considered.
   - If $s = (n_f \cdot i), i \in \mathbb{N} \Rightarrow n_m = 1$. That is, only one interleaver $(s \times s)$ matrix will be used;
   - Otherwise, $n_m = n_f$ square $(s \times s)$ matrices will be required.

   Going ahead, if we denote $rem(x, y)$ as the remainder of the integer division x/y, the writing matrices algorithm will be as follows:

- For the first matrix, we will set $R_1^i = \left\lfloor \frac{s}{n_f} \right\rfloor$, for $i = \{1, 2, \ldots, (n_f - rem(s, n_f))\}$. Similarly, we will set $R_1^j = \left\lfloor \frac{s}{n_f} \right\rfloor + 1$, for $j = \{(n_f - rem(s, n_f) + 1), \ldots, (n_f - 1), n_f\}$.
- If applicable, for the next $j = \{2, \ldots, n_f\}$ additional matrices, and for $i = \{2, \ldots, n_f\}$ flows, if $R_{(j-1)}^i = (\left\lfloor \frac{s}{n_f} \right\rfloor + 1)$ and $R_{(j-1)}^{(i-1)} = \left\lfloor \frac{s}{n_f} \right\rfloor$ then $R_j^i = \left\lfloor \frac{s}{n_f} \right\rfloor$ and $R_j^{(i-1)} = (\left\lfloor \frac{s}{n_f} \right\rfloor + 1)$.

As it can be checked, any burst of length less or equal to s will be scattered at the de-interleaver output. In this case, if we define $r = rem(s, n_f)$ and $d = (s-r)/n_f$, D_{max}, the maximum delay will obey the following expressions

- If $r \leq (n_f - r) \Rightarrow D_{max} = s \cdot (r \cdot (d+1) - 1 - (r-1) \cdot d)$.
- If $r > (n_f - r) \Rightarrow D_{max} = s \cdot (r \cdot (d+1) - 1 - ((r-1) \cdot d + 2 \cdot r - n_f - 1))$

For a given s, the lower maximum delay that we can obtain is achieved when $n_f = (s-1)$, and when $s/n_f = 2$ and $r = 0$. This delay corresponds to $D_{max} = s$. Therefore, the maximum tolerated s, given a flow with a maximum per packet time to live d_{max} and a period of t_f must satisfy that $s < \frac{d_{max}}{t_f}$. For the provided numerical example, in which $t_f = 20$ ms and $d_{max} = 300$ ms, it yields that $s < 15$, what is significantly less demanding compared to the upper bound of $s < 5$ for the *Type I* (s) end-to-end interleaver. The period of the proposed *Type II* (n_f, s) interleaver is equal to $\frac{s^2}{n_f}$, if $s \equiv 0 \, (mod \, n_f)$, and s^2 in the other case.

3 Quality and Intelligibility Evaluation

To evaluate our VoIP interleavers we plan to use a high level criterion. In noise free conditions, Automatic Speech Recognition (ASR) rate is highly correlated to human intelligibility [7]. Based on that, we propose to use this score as the performance measure. We feel that this methodology should be definitively considered to evaluate any VoIP service enhancement. Compared to MOS subjective tests, ASR has lower cost and is more reproducible. In addition, in terms of the intelligibility perceived by the user, ASR rate can be more suitable than other quality measures like *PESQ* (ITU-T recommendation P.862) or the *E-model* [8].

Speech recognizer's performance is measured in terms of the Word Error-Rate (WER), defined by:

$$WER = \frac{n_i + n_s + n_d}{n_t} \times 100 \tag{1}$$

where n_s is the number of substituted words, n_i is the number of spurious words inserted, n_d is the number of deleted words and, n_t is the overall number of words. Prior to the count of substitution, deletion and insertion errors, dynamic programming is used to align the recognized sentence with its correct transcription.

4 Experimental Results

Experimental results are provided by means of simulation. A simple scenario is set. n_f periodic flows arrive into the active router with period equal to $t_f = 20$ ms. For the *Type I* (s) case, just one flow ($n_f = 1$) is considered. Ideally, we assume no switching or any other routing delay.

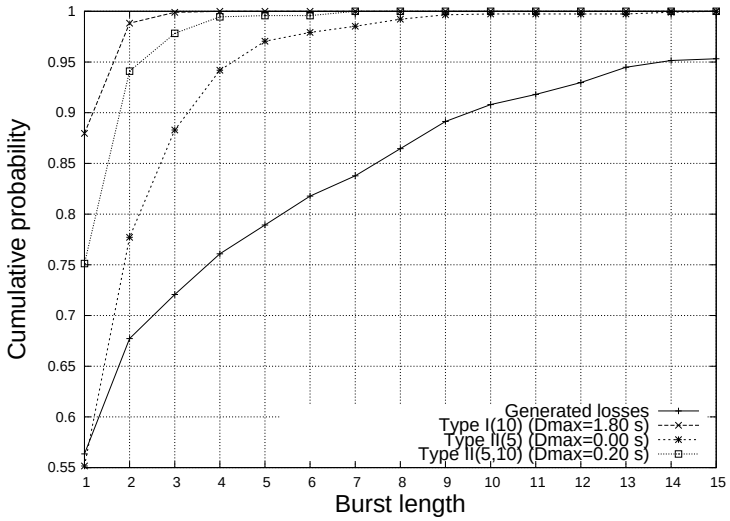

Fig. 1. Bursts length CDF

We adopt a single error model. It is based on a Markov chain trained with collected traces described in [9]. The bursts length CDF of the trace obtained from the trained model is shown in Fig. 1. The overall probability of loss is 8.2%. In the same figure, as an illustrative example ($n_f = 5$ and $s = 10$), we plot the bursts length CDFs obtained by using the simulated interleavers. Note that $TypeI(10)$ is not practically applicable, $TypeII(5)$ although does not introduce extra delay, it has less scattering capabilities and, finally, $TypeII(5, 10)$ exhibits a balance between the introduced delay and the loss isolation capacity.

To enhance the quality of the received flow, before ASR evaluation, whenever a loss packet is detected, the previous received packet will be artificially repeated. For ASR evaluation we use the connected digit Project Aurora 2 database [10]. After transmission, in order to reduce its inherent variability the speech signal is processed. A feature extractor segments the received speech signal in overlapped frames of 25 ms every 10 ms. Each speech frame is represented by a feature vector containing 13 Mel Frequency Cepstrum Coefficients and the log-Energy. Finally, the feature vectors are extended with their first and second derivatives.

The speech recognizer is based on Hidden Markov Models (HMM). We use eleven 16-state continuous HMM word models, (plus silence and pause, that have

3 and 1 states, respectively), with 3 gaussians per state (except silence, with 6 gaussians per state). The HMM models are trained from a set of 8440 noise-free sentences, while the out-of-train-test set comprises 4004 noise-free sentences.

In Table 1, WER and D_{max} values are summarized for $TypeI(s)$, $TypeII(n_f)$ and $TypeII(n_f, s)$ interleavers with different n_f and s. For a given n_f, the s value was chosen such as D_{max}, expressed in seconds, would be lower than 0.300 for $TypeII(n_f, s)$ interleaver. Note that for $TypeII(n_f)$, D_{max} is not shown because its theoretic delay is equal to 0.

It can be observed that WER performance for $TypeII(n_f)$ interleaver strongly depends on n_f. For $n_f = 2$, $TypeII(n_f, s)$ can reduce the $TypeII(n_f)$ WER without breaking the end-to-end constraint. However, as more n_f will be available, the WER difference is less noticeable. Note that although $TypeI(s)$ outperforms both $TypeII$ interleavers it can be only used for $s < 5$, given the VoIP end-to-end delay constraint.

Table 1. WER (%) and D_{max} in seconds for the three simulated interleavers.

	$II(n_f)$	$II(n_f, s)$			$I(s)$				$II(n_f)$	$II(n_f, s)$			$I(s)$		
n_f	WER	s	WER	D_{max}	s	WER	D_{max}	n_f	WER	s	WER	D_{max}	s	WER	D_{max}
2	5.401	3	4.610	0.060	3	3.080	0.120	6	2.420	12	1.449	0.240	12	1.263	2.640
		5	2.892	0.200	5	1.595	0.400	7	1.968	14	1.381	0.280	14	1.340	3.640
		6	2.255	0.240	6	1.540	0.600	8	1.819	9	1.626	0.180	9	1.304	1.440
3	4.333	7	1.831	0.280	7	1.307	0.840	9	1.613	10	1.526	0.200	10	1.285	1.800
4	3.419	8	1.848	0.160	8	1.289	1.120	10	1.611	11	1.533	0.220	11	1.282	2.200
5	2.875	10	1.567	0.200	10	1.325	1.800	12	1.534	13	1.455	0.260	13	1.322	3.120

Additionally, note that the $TypeII(4, 8)$ maximum delay (160 ms) is lower than the $TypeII(3, 7)$ delay (280 ms), although the s value is greater. This is due to the peculiarity of the interleaver, which results in non linear delay dependence with the number of flows and the burst length considered.

Summing up, both the Type II interleavers diminish the packet interleaving delay. Although Type II (n_f) is designed to work properly when $n_f \geq s$, it can be suited when $n_f \approx s$, without introducing any additional delay. Compared to Type II (n_f), the Type II (n_f, s) interleaver improves the VoIP intelligibility. It scatters a high percentage of losses patterns, and reduces the maximum length of the bursts at the de-interleaver output. Furthermore, although $TypeII(n_f, s)$ introduces some additional delay, it can be used under conditions that $TypeII(n_f)$ does not tolerate (long burst length and low number of different flows).

On a separate note, given the processing capabilities of the active router, it could be always possible to select which interleaver algorithm to use ($TypeII(n_f)$ or $TypeII(n_f, s)$). In this case, the network dynamics and the number of available flows should be taken into account. As a rule of thumb, we would suggest the consideration of the $TypeII(n_f)$ interleaver instead of $TypeII(n_f, s)$ whenever $n_f \approx s$.

5 Conclusion

In this paper the block interleaving problem for audio applications is revisited. To increase the final audio quality we aim to scatter long bursts of packet losses. We propose a new VoIP interleaver algorithm which not only diminish the per packet delay, but also allows its use under conditions where end-to-end approaches are unfeasible. Our algorithm interleaves packets from different flows. To work properly, the interleaver must be placed in a common node before the path where losses are expected to occur. We show that the resulting speech intelligibility is maximized, especially when the number of available flows is small.

In this work, because of its reproducibility and low cost, we have considered automatic speech recognition in order to assess the intelligibility improvements of the proposed VoIP active service. This procedure can be extended to evaluate any other VoIP enhancement. As future work, to establish a mapping function for human to machine recognition rate remains. Similarly, the mapping functions between recognition rate and MOS score should be studied as well. By using these mapping functions, enhanced VoIP active services can be envisaged.

References

[1] Liang, Y.J.; Farber, N.; Girod, B.: Adaptive playout scheduling and loss concealment for voice communication over IP networks. IEEE Transactions on Multimedia, vol. 5, no. 4, Dec. 2003 Page(s): 532-543.

[2] Towsley, D.; Kurose, J.; Pingali, S.: A comparison of sender-initiated and receiver-initiated reliable multicast protocols. IEEE Journal on Selected Areas in Communications, vol. 15, no. 3, 1997 Page(s):398-406.

[3] Tennenhouse, D.L.; Smith, J.M.; Sincoskie, W.D.; Wetherall, D.J.; Minden, G.J.: A survey of active network research. IEEE Communications Magazine, vol. 35, no. 1, 1997 Page(s): 80-86.

[4] Ott, D.E.; Sparks, T.; Mayer-Patel, K.: Aggregate Congestion Control for Distributed Multimedia Applications. IEEE INFOCOM 2004. Volume 1, March 2004, Page(s): 13-23.

[5] Ramos-Muñoz, J.J.; Lopez-Soler, J.M.: Low Delay Multiflow Block Interleavers for Real-Time Audio Streaming. Lecture Notes in Computer Science, vol. 3420, Jan 2005, Page(s): 909-916.

[6] Kenneth Andrews, C.H.; Kozen, D.: A theory of interleavers. Technical Report 97-1634, Computer Science Department, Cornell University, 1997.

[7] Jiang, W.; Schulzrinne, H.: Speech recognition performance as an effective perceived quality predictor. Tenth IEEE International Workshop on Quality of Service, May 2002, Page(s): 269-275

[8] Cole, R.G.; Rosenbluth. J.H.: Voice Over IP Performance Monitoring. SIGCOMM Comput. Commun. Rev., vol. 31 (2), 2001, Page(s): 9-24.

[9] Yajnik, M.; Kurose, J.; Towsley, D.: Packet loss correlation in the MBone multicast network experimental measurements and markov chain models. Tech. Rep. UM-CS-1995-115, 1995.

[10] Hirsch, H.G.; Pearce, D.: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition, ISCA ITRW ASR2000, France, 2000.