A Dynamic FEC for Improved Robustness of CELP-Based Codec

Nadir Benamirouche[1], Bachir Boudraa[2], Angel M. Gomez[3],
José L. Pérez-Córdoba[3](✉), and Iván López-Espejo[3]

[1] Laboratoire de Génie Electrique, Faculté de Technologie,
Université de Bejaia, 06000 Bejaia, Algeria
benam_nadir@yahoo.fr

[2] Faculty of Electronics and Computer Science,
University of S.T.H.B, Algiers, Algeria
b.boudraa@yahoo.fr

[3] Department of Signal Theory, Networking and Communications,
University of Granada, 18071 Granada, Spain
{amgg,jlpc,iloes}@ugr.es

Abstract. The strong interframe dependency present in Code Excited Linear Prediction (CELP) codecs renders the decoder very vulnerable when the Adaptive Codebook (ACB) is desynchronized. Hence, errors affect not only the concealed frame but also all the subsequent frames. In this paper, we have developed a Forward Error Correction (FEC)-based technique which relies on energy constraint to determine frame onset which will be considered for sending the FEC information. The extra information contains an optimized FEC pulse excitation which models the contribution of the ACB to offer a resynchronization procedure at the decoder. In fact, under the energy constraint the number of Fixed Codebook (FCB) pulses can be reduced in order to be exploited by the FEC intervention. In return, the error propagation is considerably prevented with no overload of added-pulses. Furthermore, the proposed method greatly improves the CELP-based codec robustness to packet losses with no increase in coder storage capacity.

Keywords: Speech coding · VoIP · Forward error correction · Lossy packet networks · Error propagation · ACB resynchronization

1 Introduction

Media streaming services such as Voice over Internet Protocol (VoIP) is an emerging technology which has become a key driver in the evolution of voice communications. Unfortunately, the quality of service (QoS) of VoIP does not yet provide toll-quality voice equivalent to that offered by the traditional public switched telephone network [1]. Indeed, another critical issue for media streaming applications such as VoIP, is its vulnerability to end-to-end performance [2,3]. Some packets may be delayed or lost due to network congestion. Hence, the missing packets have to be regenerated at the decoder side using packet

A. Abad et al. (Eds.): IberSPEECH 2016, LNAI 10077, pp. 14–23, 2016.
DOI: 10.1007/978-3-319-49169-1_2

loss concealment techniques. In Code Excited Linear Prediction (CELP)-based codecs [4] a long-term predictor (LTP) is used to encode the excitation signal through its past samples. Since such speech parameters are not efficiently estimated by the concealment approach, it is reported that this mismatch on the obtained excitation introduces error propagation through the properly received frames [6]. A wide variety of error concealment techniques have been proposed as solutions for the above problems in order to mitigate the effect of lost frames, especially in the context of voice over packet networks [1–7]. These alternative solutions are based on considering some side information sent as extra bit-rate. In the same direction, recent approaches are focused on limiting the inter-frame dependencies using forward error correction (FEC) [8], where additional information are used to reinitialize the decoder in presence of packet loss. This redundancy provides an alternative representation of the previous excitation samples to prevent error propagation [7].

In this paper, we propose a decoder resynchronization method based on dynamically adding FEC information to improve the robustness of CELP-based codecs. As a solution to the error propagation problem, the proposed method sends side information replacing the Adaptive Codebook (ACB) contribution with a set of pulses as memoryless codebook. The developed FEC-based technique relies on energy ratio constraint to determine voiced subframes (frame onset) which side information will be considered for. Through the proposed method, a higher reduction in bit-rate is achieved when the extra information is sent for only the first two subframes (SF_0 and SF_1) of a voiced frame, since the pitch component of the subsequent subframes (SF_2 and SF_3) can be estimated using the previous two subframes. The proposed method is a subframe-based technique which uses the Least Square Error (LSE) criterion over the synthesized speech domain to optimize the FEC pulses. Hence, at the decoder side when the previous frame is erased, this set of pulses is added to the Fixed Codebook (FCB) vector to form the total excitation. The synthesized speech signal is finally obtained as the LP filter response to this excitation. Through this approach the ACB resynchronization pulse search does not significantly increase the overall complexity of the CELP-based codec.

This paper is organized as follows. In Sect. 2, we present the ACB resynchronization approach using reduced-FCB pulse compensation and the applied criterion for its optimization. Subsequently, in Sect. 3, we describe the experimental framework applied to simulate lossy packet channels, the used speech database and objective quality measure to assess the quality of the proposed method. In Sect. 4, we discuss the effectiveness of the proposed method, where the obtained results under the FEC method intervention are shown. Finally, conclusions of this work are summarized in Sect. 5.

2 ACB Resynchronization Approach Using Reduced-FCB Pulse Compensation

Under the CELP model, a segment of synthesized speech for each subframe is obtained by filtering an error signal (1), by means of a short-term linear prediction (LP) filter, *1/A(z)*. After removing the contribution of the LP filter memory, the new version of the error signal, $\widehat{e}(n)$, can be expressed as follows,

$$\begin{aligned}\widehat{e}(n) &= x(n) - \widehat{x}(n),\\ &= x(n) - \sum_{j=0}^{N-1} h(j) \cdot \widehat{e}(n-j),\end{aligned} \tag{1}$$

where *x(n)* indicates the target signal once the contribution of the LP filter memory has been removed, $\widehat{x}(n)$ is the synthesized one, *h(j)* the impulse response of the LP filter and N is the subframe length. Similar for most CELP-based codecs, the excitation signal given in (1) consists of two components, the FCB (Fixed Codebook) excitation $e_f(n)$, and ACB (Adaptive Codebook) excitation $e_a(n)$ also known as Long-Term Prediction (LTP) contribution. Formally, the total excitation signal, $\widehat{e}(n)$, is obtained as

$$\begin{aligned}\widehat{e}(n) &= \sum_{j=-(l-1)/2}^{(l+1)/2} b(j) \cdot e(n-(T+j)) + g_f \cdot e_f(n)\\ &= e_a(n) + g_f \cdot e_f(n),\end{aligned} \tag{2}$$

where T, $b(j)$ and g_f are the pitch lag, LTP filter and fixed vector gain respectively, and l is the order of the LTP filter. The goal of the innovative codebook contribution $e_f(n)$ is to model the residual signal remaining after removing the long-term redundancy.

2.1 Frame Onset Detection

In order to determine the onset frames, an energy ratio is continuously computed for subframes SF_0 and SF_1. This energy ratio is based on the ACB excitation energy and the target signal energy. Let us suppose that Q is the quotient of ACB contribution energy noted E_{ACB}, over the target signal energy noted E_x, that is

$$Q = \frac{E_{ACB}}{E_x} \tag{3}$$

where, $E_{ACB} = \sum_{n=0}^{N-1} e_a^2(n)$, and, $E_x = \sum_{n=0}^{N-1} x^2(n)$. Then, the energy ratio, Q, is compared to a predefined energy threshold, Q_c. If Q is greater than or equal to Q_c, implies the corresponding subframe is judged as important (voiced), otherwise, the subframe is not important (unvoiced).

To set the value of Q_c, we performed multiple tests under different values of the energy ratio threshold in the range of [0.1, 1] with respect to three lossy channels of 6 %, 13 % and 18 % loss rate. Figure 1 shows the obtained PESQ (Perceptual Evaluation of Speech Quality) scores [9] as a function of the energy ratio threshold Q_c with respect to different channel erasure conditions. In regard to the depicted curves in Fig. 1, we can notice that the coding robustness is

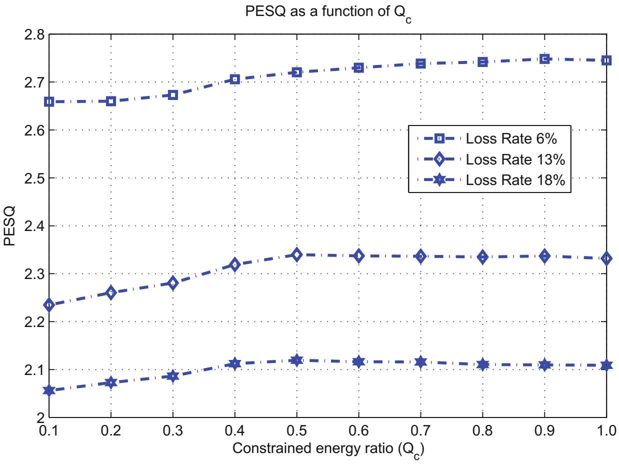

Fig. 1. Obtained PESQ scores of FEC method by varying the energy ratio threshold with respect to different channel erasure conditions and without any reduction of FCB pulses.

significantly improved around Q_c equal 0.5. Obviously, under this constraint $(Q \geq Q_c)$, more than a half of synthesized speech signal energy is within the ACB contribution, which means the FCB contribution has lower energy contribution from the total excitation energy.

$$\frac{E_{ACB}}{E_x} \geq 0.5, \Rightarrow \; E_{ACB} \geq \frac{E_x}{2}. \tag{4}$$

2.2 Applied Least Square Error Criterion

The proposed method uses the Least Square Error (LSE) criterion in order to provide the minimum error between the synthesized speech signal $\widehat{x}(n)$ and the original speech signal $x(n)$, where *h(n)* is the impulse response of the LP filter and N is the subframe length. For this purpose, the error is given by

$$\varDelta = \sum_{n=0}^{N-1}(x(n) - \widehat{x}(n))^2 = \sum_{n=0}^{N-1}(x(n) - h(n) * \widehat{e}(n))^2. \tag{5}$$

To take into account the human auditory perception, the error signal is commonly weighted by a perceptual filter, *W(z)*, so that

$$\begin{aligned}\varDelta_w &= \sum_{n=0}^{N-1}(w(n) * (x(n) - \widehat{x}(n)))^2, \\ &= \sum_{n=0}^{N-1}(x_w(n) - h_w(n) * \widehat{e}(n))^2. \end{aligned} \tag{6}$$

The CELP excitation $\hat{e}(n)$ can be considered as a summation of two signals, namely, the zero state and the zero input excitation. On one hand, the zero state excitation is computed by considering that the samples before the current frame are zero (i.e., no samples on the memory). On the other hand, the zero input excitation is obtained by considering that the fixed vector is zero for the current frame (i.e. the input is null). To resolve the LSE optimization problem, we can redefine the excitation signal as the resulting signal of the filter *P(z)*,

$$P(z) = \frac{g_f}{1 - \sum\limits_{j=-(l-1)/2}^{(l-1)/2} b(j) z^{-(T+j)}}. \tag{7}$$

Therefore, the excitation signal can be obtained as the sum of the zero state $\widehat{e}_{zs}(n)$ and zero input $\widehat{e}_{zi}(n)$ responses from the filter *P(z)*. Under this assumption, the quadratic error to be minimized can be expressed as

$$\varDelta_w = \sum_{n=0}^{N-1} (x_w(n) - h_w(n) * (\widehat{e}_{zs}(n) + \widehat{e}_{zi}(n)))^2. \tag{8}$$

2.3 FEC Size Reduction

Firstly, in order to reduce the binary payload introduced by coding the FEC pulses, the second part of (2) is modified, so that the number of FCB pulses, M, is reduced to $(M - \alpha)$ and the new expression for the fixed codebook is

$$e_f(n) = \sum_{i=1}^{M-\alpha} g_f \cdot \delta(n - m_i), \tag{9}$$

where $(1 \leq \alpha \leq K)$ is the number of the subtracted pulses from FCB codebook, $K = 3$, m_i indicates the pulse position 'm' with index 'i' in the FCB excitation, with $i = 1, ..., M - \alpha$ and M is a legacy number of FCB pulses which is greater than α. Secondly, the contribution of the reduced FCB pulse vector $e_f(n)$ in (9), can be removed from the optimization to deal with its introduced complexity. In this new context, the quadratic error to be minimized can be expressed as

$$\varDelta_w = \sum_{n=0}^{N-1} (x_w(n) - x_{zs}(n) - h_w(n) * \widehat{e}_{zi}(n))^2, \tag{10}$$

where $\widehat{e}_{zs}(n)$ is the zero state contribution introduced by the reduced-FCB and $x_{zs}(n)$ is the LP response to the fixed vector excitation. As a result, the excitation signal can be defined as a recursion of only adaptive contributions which finally depends on a set of initial pulses placed on the ACB memory. Since the zero-state does not depend on previous samples by definition, we can remove it from the optimization so the final square error to be minimized is given by

$$\varDelta_w = \sum_{n=0}^{N-1} (t(n) - h_w(j) * \widehat{e}_{zi}(n))^2. \tag{11}$$

Hence, the problem is now simplified to optimize the Least Square Error (LSE) criterion between the modified target signal *t(n)* and the weighted-synthesis filter response $h_w(n)$, excited with an optimized FEC pulse excitation signal, (seen as a zero input response). Therefore, the zero input excitation $\widehat{e}_{zi}(n)$ that we are searching for is,

$$\widehat{e}_{zi}(n) = \sum_{k=1}^{K} g_{m_k} \delta(n - m_k), \tag{12}$$

where g_{m_k} is the amplitude of the pulse, $\delta(n - m_k)$ is the unit pulse, m its position, k is the index of each pulse to be set, while *(K= 1, 2, 3)*. Thus, the pulse positions and amplitudes are optimized using the MP-MLQ algorithm [4].

2.4 The Proposed ACB Resynchronization Scheme

Figure 2 shows the modified encoder scheme for FEC pulse optimization. The dashed boxes in Fig. 2 represents the added functions to perform this optimization. It must be pointed out that the resynchronization parameter search is applied only if the imposed condition on the energy ratio is satisfied. Subsequently, at the decoder side, the ACB resynchronization procedure takes place once the previous frame is erased. In this case, the added FEC pulses are used to replace the desynchronized ACB codebook then the generated total excitation $\hat{e}(n)$ is used to resynchronize the ACB memory.

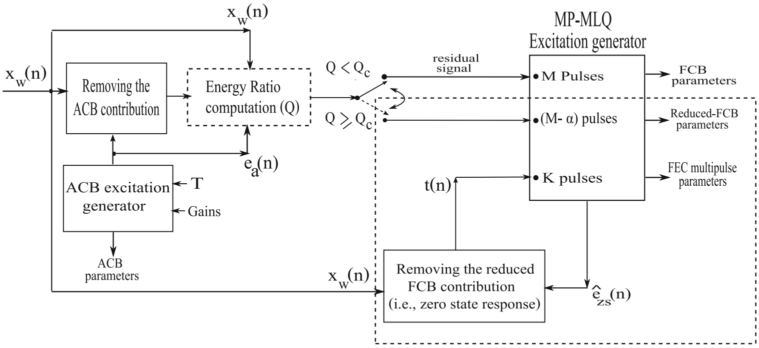

Fig. 2. Modified encoder scheme for FEC side information optimization.

3 Experimental Framework

We have used the standard codec G.723.1 [4] as a CELP codec for carrying the experiments to check the improvements of our proposed technique. Likewise, for performance evaluation, we have considered an objective test performed by means of the ITU Perceptual Evaluation of Speech Quality standard (PESQ) [9]. In order to provide an objective quality measure, PESQ is applied over a subset of the well-known TIMIT database which contains broadband recordings from 630 speakers of eight major dialects of American English [10,11]. To this end, testing and training utterances from the TIMIT database are down-sampled to 8 kHz and their lengths artificially extended to approximately 14 s. For each used utterance, the PESQ algorithm provides a score within a range from −0.5 (bad) to 4.5 (excellent). In order to obtain an overall score for each channel condition, the score of each sentence is weighted by its relative length.

During these simulations, packet loss rates of 6, 8, 10, 13, 16, 18, 20, 21 and 23 % were generated by the Gilbert-Elliot model defined in [12]. Under these frame loss channels, the burst of consecutive frame losses is varying from 1 up to 3 frames. Likewise, for our study the larger size of burst losses cannot be considered since the used CELP-based codec design does not support more than 3 successive frame losses [4] and the error propagation effects are usually noticed.

4 Experimental Results

To assess the efficiency of sending side information only for the first two subframes instead of all the four subframes Fig. 3 illustrates a comparison between PESQ scores obtained from both cases, relative to the variation of reduced FCB pulse $(\mathrm{M} - \alpha)$ under loss rate conditions (6 %, 13 %, 18 % and 23 %) with a constrained energy ratio Q_c at 0.5. We can notice from the obtained figures in Fig. 3 that the PESQ scores for 6 % and 13 % of loss are similar either, for first two important SFs or four important SFs, even if the number of subtracted pulses α from FCB is increased from 1 to 3. This returns relatively to the loss rates, 6 % and 13 % which render the FEC method intervention very restricted. In contrast, under 18 % and 23 % of loss, we recorded a noticeable improvement of PESQ quality when we consider only the first two SFs from each voiced frame relative to four SFs per voiced frame. Since it has been noticed that most of pitch-lag values are less than or equal to 120, due to this feature, the obtained

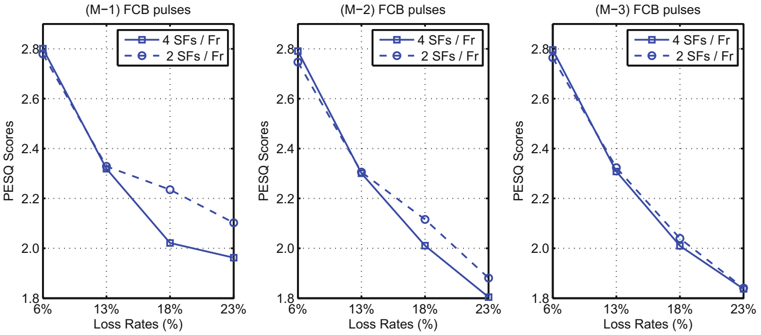

Fig. 3. PESQ Scores as a function of loss rate and reduced FCB pulse number $(\mathrm{M} - \alpha)$ for 2 important SFs and 4 important SFs per frame, respectively.

results confirm the efficiency of our choice to send extra information for only first two subframes. Accordingly, the obtained results confirm once again the effectiveness of ACB in modeling the glottal pulses when this latter is updated. In other words, it is more efficient to use the resynchronized ACB for SF_2 and SF_3 instead of FEC pulses since it has been updated by the recovered first two subframes. In return, in case of $(\mathrm{M}-3)$ for 18 % and 23 % of loss rate, the recorded equality in PESQ values between 2 SFs and 4 SFs, returns to the limits of the established tradeoff between reduced-FCB and pulse compensation efficiency to model the ACB. This feature led us to restrict the variation of α up to K, which means that the targeted performance of the proposed method is not guaranteed for a higher value of α. To emphasize on the proposed method robustness, a wide range of channel conditions based on Gilbert-Elliot model is tested, including frame erasure ratios of 6 %, 8 %, 10 %, 13 %, 16 %, 20 %, 21 % and 23 % in order to exhaustively evaluate the efficiency of this latter against error propagation. Therefore, Fig. 4 shows the obtained PESQ results related to the performance of the proposed method by varying the number of subtracted pulses α from FCB codebook. In addition, the results from a complete LTP restoration (Restore memory) and those obtained from legacy standard G723.1 codec [4] are also shown in this figure. As expected, reducing FCB pulse $(\mathrm{M}-\alpha)$ for important subframes does not strongly affect the quality of synthesized speech signal. This can easily be explained by the fact that most of the excitation energy is within the ACB contribution. Indeed, as we can see in Fig. 4 under high loss rates, there is a great improvement of PESQ quality offered by the proposed FEC method while it is approaching the quality of complete memory restoration compared with the legacy G723.1 codec.

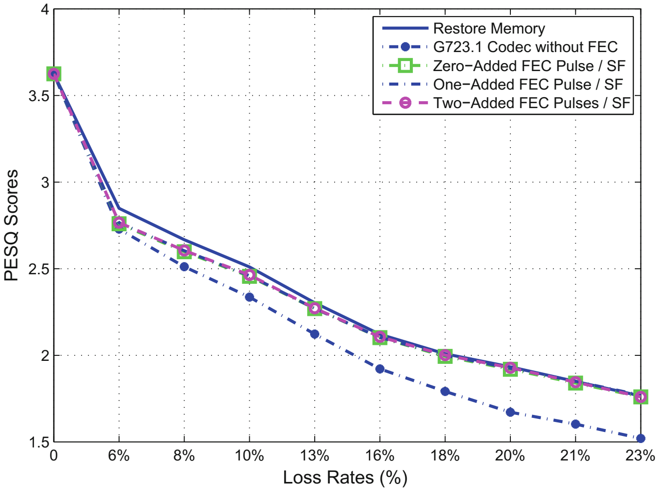

Fig. 4. PESQ scores obtained from legacy G723.1 codec, G723.1 with complete ACB restoration and from FEC based technique intervention relative to two, one or zero added-pulse under different lossy channel conditions.

Table 1. PESQ results obtained by FEC multi-pulse relative to the number of added-pulses at multiple loss rate conditions.

Added pulses	Packet loss ratio								Avg.
	6 %	8 %	10 %	13 %	16 %	18 %	20 %	23 %	
0 P.	2.7596	2.5977	2.4561	2.2693	2.1018	1.9937	1.9194	1.7601	2.2322
1 P.	2.7651	2.6041	2.4598	2.2704	2.1035	1.9997	1.9251	1.7696	2.2372
2 P.	2.7664	2.6052	2.4638	2.2720	2.1062	2.0011	1.9278	1.7602	2.2378

Table 1, lists the obtained PESQ results corresponding to different cases of reduced-FCB (M−1, M−2 and M−3), with respect to the packet loss conditions. The results show a very small variation of PESQ averages although the value of α, is increased from 1 to 3. It has also been proved that reducing FCB pulses under certain limits does not really affect the quality of the proposed method. Thus, the subtracted pulses from FCB are replaced by FEC pulse insertion which offers the opportunity to improve the decoder robustness with no overload of extra pulses, particularly when α is equal to 3.

5 Conclusions

In this paper, we have proposed a FEC method to offer a resynchronization step for CELP-based G732.1 codec which relies on dynamically adding side information. In particular, a constrained energy ratio is applied to determine the important subframes in each voiced frame. The proposed method consists of constrained optimization of FEC pulse excitation at the encoder and a resynchronization procedure at the decoder. Therefore, when the resynchronization procedure is performed with respect to the first two subframes, the LTP parameters of the subsequent subframes can be easily predicted. Accordingly, the aim behind the proposed method is to achieve a reduced computational complexity with a resulting bit-rate that would be much lower compared to the increase that can be obtained by sending the extra information every frame. Furthermore, the objective quality tests under frame erasure conditions have shown the suitability of the proposed technique. Finally, the speech quality evaluation confirmed that the pulse compensation of ACB, introduces a very small bit-rate increase and achieves a noticeable improvement of objective quality which approaches a complete ACB memory restoration. Also, other approaches may benefit from this method and can contribute to a better robustness against error propagation.

References

1. Toral, C.H., Pathan, A.K., Ramirez, P.J.C.: Accurate modeling of VoIP traffic QoS parameters in current and future networks with multifractal and Markov models. Math. Comput. Model. **57**(11), 2832–2845 (2013)

2. Bhebhe, L., Parkkali, R.: VoIP performance over HSPA with different VoIP clients. Wirel. Pers. Commun. **58**(3), 613–626 (2011)
3. Kim, B.H., Kim, H.G., Jeong, J., Kim, J.Y.: VoIP receiver-based adaptive playout scheduling and packet loss concealment technique. IEEE Trans. Consum. Electron. **59**(1), 250–258 (2013)
4. ITU Rec.: G.723.1, dual rate speech coder for multimedia communication transmitting at 5.3kbit/s and 6.3kbit/s (1996)
5. Oh, S.M., Kim, J.H.: Application-aware retransmission design for VoIP services in BWA networks. In: 14th International Conference on Advanced Communication Technology (ICACT), pp. 122–131. IEEE (2012)
6. Gomez, A.M., Carmona, J.L., Peinado, A., Sanchez, V.: A multipulse-based forward error correction technique for robust CELP-coded speech transmission over erasure channels. IEEE Trans. Audio Speech Lang. Process. **18**(6), 1258–1268 (2010)
7. Gomez, A.M., Carmona, J.L., Peinado, A., Sanchez, V.: One-pulse FEC coding for robust CELP-coded speech transmission over erasure channels. IEEE Trans. Multimed. **13**(5), 894–904 (2011)
8. Ehara, H., Yoshida, K.: Decoder initializing technique for improving frame-erasure resilience of a CELP speech codec. IEEE Trans. Multimed. **10**(3), 549–553 (2008)
9. Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T P.862 Recommendation (2001)
10. Lamel, L., Kassel, R., Seneff, S.: Speech database development: design and analysis of the acoustic-phonetic corpus. In: Proceedings of Speech Recognition Workshop (DARPA), pp. 100–110 (1986)
11. Garofolo, J.S.: The Structure and Format of the DARPA TIMIT, CD-ROM Prototype, Documentation of DARPA TIMIT
12. Jiang, W., Schulzrinne, H.: Modeling of packet loss and delay and their effect on real-time multimedia service quality. In: Proceedings of NOSSDAV 2000 (2000)

http://www.springer.com/978-3-319-49168-4

Advances in Speech and Language Technologies for
Iberian Languages
Third International Conference, IberSPEECH 2016,
Lisbon, Portugal, November 23-25, 2016, Proceedings
Abad, A.; Ortega, A.; Teixeira, A.J.d.S.; Garcia Mateo, C.;
Martínez Hinarejos, C.D.; Perdigão, F.; Batista, F.;
Mamede, N. (Eds.)
2016, XV, 288 p. 53 illus., Softcover

ISBN: 978-3-319-49168-4