Coded-speech recognition over IP networks

Ph.D. José L. Carmona
Advisors: Antonio M. Peinado, José L. Pérez-Córdoba

Dpto. Teoría de la Señal, Telemática y Comunicaciones, University of Granada
{maqueda, amp, jlpc} @ ugr.es

Abstract

In this Ph.D. dissertation the influence of packet losses on speech recognition is analyzed and different solutions to prevent, reduce and conceal their effects are developed. The performance of remote speech recognition will be subject to the robustness of the speech coding scheme used. Conventional speech codecs achieve to reduce the bit-rate by making use of predictive techniques that exploit temporal speech correlations. Thus, to decode a frame, a correct decoding of the previous ones is required. However, this inter-frame dependency reduces considerably the robustness against packet losses because it originates an error propagation in addition to the associated information loss. Furthermore, speech decoders integrate their own packet loss concealment algorithms, which are based on perceptual considerations that are unsuitable for speech recognition. In order to combat these degradations, we propose a set of mechanisms that can be divided into sender-driven and receiver-based techniques.

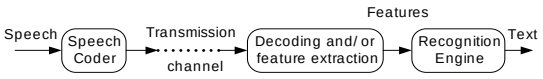
Index Terms: Network speech recognition, robust speech recognition, packet loss concealment.

1. Nature of the problem

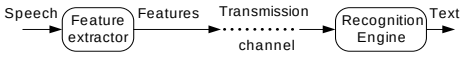
Thanks to the convergence of wireless technologies, access to information on the move is nowadays a technological reality on the increase. Nevertheless, mobile phone constraints, such as the lack of keypad for size reasons, hinder the access to remote services. Oral interaction with such services arises as a new faster and more natural means of access to information with the help of automatic speech recognition that offers an interactive service and fast access to information, what benefits the user, doing without the assistance operator at the other side, what benefits the provider. Unfortunately, there are several problems to install an automatic speech recognition subsystem into mobile terminals. It is mainly their size restrictions what limits the computation capacity and, therefore, the recogniser power and flexibility. The possibility of remote speech recognition, i.e. outside the terminal, emerged to overcome these obstacles.

Remote Speech Recognition (RSR) allows to circumvent these hardware constraints by moving the most complex computational tasks of speech recognition to a remote server. Moreover, the structure of a remote recognition system is well suited for the IP model, since it is the provider that implements the recogniser depending on its needs. Thus, the provider can incorporate new services adapted to the present needs of users. Under this point of view, low cost terminals with limited features are connected to powerful remote computers that carry out more complex tasks for them, what leads to an optimum use of centralised resources. As shown in Fig. 1, there are two possibilities for the implementation of an RSR system [1]:

1. Network Speech Recognition (NSR). In this approach, the whole recognition system resides in the network. Thus, the client sends the speech signal, employing a conventional speech codec, to the server where recognition is carried out.
2. Distributed Speech Recognition (DSR). The client includes a local front-end that processes the speech signal in order to obtain the specific features used by the remote server (back-end) to perform recognition.



(a) NSR



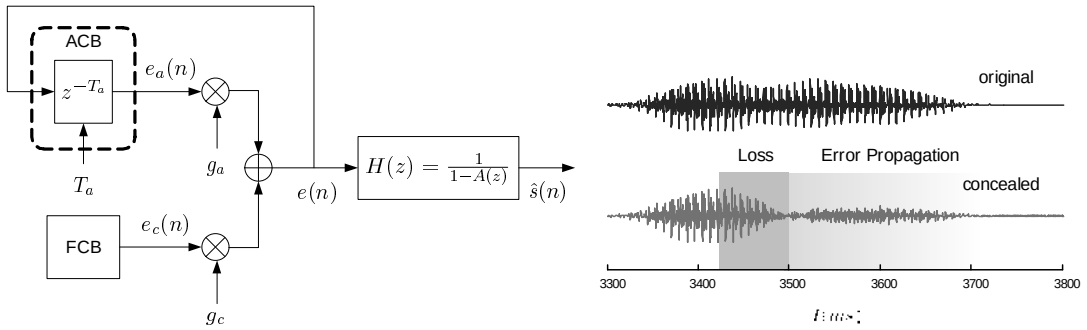
(b) DSR

Figure 1: Different architectures for the implementation of a Remote Speech Recognition (RSR) system. a) Network Speech Recognition (NSR); b) Distributed Speech Recognition (DSR).

In DSR, the feature extractor is applied directly to the speech signal to obtain a low dimensional representation with less redundant information. Although during the last years several standards have been issued, the lack of DSR codecs in the existing devices supposes a barrier for its deployment.

On the other hand, the most direct implementation of an RSR system is the speech transmission to the extreme server where the recognition task is performed, i.e., the NSR architecture. In this case, RSR is considered a value-added service of VoIP, since the coded speech is not transmitted to establish a call session, but to have access to a particular service. The main advantage of this type of application is the use of emerging IP platforms, without modifying in any respect the client terminal. However, the application has at the same time some disadvantages, since the loss information that speech coding involves may affect performance. There are also some implicit problems in remote recognition. Among them, two of the most outstanding ones are acoustic noise (the acoustic context of the terminal may vary) and degradations introduced by the communication channel. This dissertation is focused on the second one, since IP networks design, which offer a best-effort service, does not guarantee to meet real time requirements (delay and jitter) nor reliability requirements of multimedia flow transmission (packet losses).

This thesis [2] aims to develop a set of mechanisms to improve the performance of NSR systems considering the degradation of the current IP networks. In particular, such mechanisms should overcome the loss of VoIP packets over the network, optimising the performance of recognition tasks. In this sense, two types of measures can be adopted: sender-driven or



(a) (b)

Figure 2: Impact of a packet loss on a CELP-based codec: (a) Decoder structure; (b) Speech synthesis applying the packet loss concealment included in the speec codec.

receiver-based. The first type of these measures attempts to make speech coding schemes more robust against packet loss effects. In this case, modifications in the coding scheme will be applied taking into account a double objective: improve both the subjective perceptual quality of coding systems and speech recognition. The second type of measures will develop loss concealment techniques for coded speech recognition. In this case, these techniques only require to modify the server structure, and therefore they focus on the improvement of the recogniser performance disregarding perceptual aspects.

The rest of this extended abstract is organized as follows. In Section II we described the methodology that we followed to develop this Ph.D. thesis. Section III is devoted to describe the experimental framework and the most relevant results. In Section IV we summarize our conclusions. Finally, Section V collects the most important references and publications derived from this Ph.D. dissertation.

2. Methodology

The development of this research has been carried out by means of a methodology based on three stages:

1. Preliminary study of basic NSR systems. In this part we analyzed the performance degradation caused by packet losses in the speech decoding process.
2. Proposals for robust speech encoding. At this stage we took as initial hypothesis that if the proposed techniques achieve strengthen coding schemes from a perceptual point of view, then also increase the performance of recognition systems. Finally, we verified this hypothesis by carrying out speech recognition tests.
3. Proposals for packet loss concealment algorithm at the receiver. Here, we proposed different algorithms for the concealment of those degradations originated by packet losses at the receiver side.

2.1. Speech decoding in presence of packet losses

Most current speech codecs are based on the CELP (Code Excited Linear Prediction) paradigm, since it provides a high-quality synthesis at a low bit-rate. In particular, the speech synthesis is carried out by filtering the excitation signal $e(n)$ through the LPC (Linear Predictive Coding) filter $H(z)$. The general structure of a CELP decoder is shown in Fig. 2a. As we can see, the excitation signal is produced as the sum of the signals $e_p(n)$, obtained from an adaptive codebook (ACB), and $e_c(n)$, obtained from a fixed codebook (FCB), weighted by their corresponding gains g_p and g_c. The fixed codebook contains a number of innovation sequences, whereas the adaptive codebook models the long-term correlation of the excitation signal (related to the pitch period). For this reason, the adaptive codebook is dynamically built from the previous excitation samples by means of a long-term predictor (LTP).

Unfortunately, the inter-frame dependencies in the encoding process endanger the performance in packet networks. Fig. 2b illustrates this issue. As shown, once a packet loss occurs, the predictive schemes used by the encoder prevent from obtaining the correct decoded parameters. Furthermore, even when the decoder has already the correct parameters, there exists an error propagation caused by the ACB contribution to the excitation signal [3].

2.2. Robust speech coding techniques

Sender-driven loss concealment techniques for NSR are only justifiable if they optimise both perceptual quality and speech recognition accuracy. The main advantage of the NSR architecture is that no terminal changes are required, so modifications can not be justified only by recognition improvements. For this reason, our proposals are based on modifications of the coding scheme to improve the speech synthesis and, subsequently, the recognition performance.

iLBC is a speech codec specially conceived for packet networks, such as Internet, since it was designed to combat packet losses. To achieve this goal, iLBC does not exploit the correlation between adjacent frames in the excitation encoding. Thus, iLBC removes the interframe dependencies at the cost of a higher bit-rate than other coding techniques [4].

We proposed combining iLBC and CELP schemes in order to obtain a robust performance against packet losses while reducing the bit-rate of iLBC. The idea is based on using independent (iLBC) and dependent (CELP) frames in the same way that video codecs do. Thus, in case of packet losses, the error propagation of CELP frames is limited by the iLBC frames (key frames), which act as firewalls. At the same time, the insertion of CELP frames reduces the bit-rate [5].

Our second approach is also oriented to remove the error propagation caused by packet losses. In this sense, a FEC (Forward Error Correction) code consisting of the previous excitation samples for every frame will remove the possible error propagation. Of course, this completely alienates the CELP

ASR Architecture	*Bit-Rate (kbps)*	*Channel Conditions*					*Avg. Value*
		C0	*C1*	*C2*	*C3*	*C4*	
G.729 baseline	*8*	98.81	98.02	89.87	83.13	75.98	89.16
AMR baseline	*12.2*	98.68	97.93	93.97	88.55	83.07	92.44
iLBC baseline	*15.2*	98.96	98.56	96.35	92.43	87.11	94.68
G.729 B-NSR	*8*	98.82	98.52	97.64	95.69	92.89	96.71
AMR B-NSR	*12.2*	98.79	98.59	97.69	95.96	93.61	96.93
iLBC B-NSR	*15.2*	98.94	98.85	98.22	96.29	93.32	97.12
DSR FE	*4.75*	99.04	99.04	98.65	97.10	94.10	97.59

Table 1: Summary of the most relevant results.

coding idea, increasing the bitrate up to unusable limits. Instead, we proposed to encode only the most representative excitation samples by means of a multipulse scheme [6, 7].

2.3. Receiver-based PLC techniques

The packet loss concealment algorithms implemented in the decoders are unsuitable for recognition tasks. Such algorithms are based on perceptual considerations that are not appropriate for recognition. Thus, when several consecutive packets are lost, decoders progressively mute, leading to an increase on the insertion errors in the recogniser (artificial silences).

Under the perspective of an NSR system, packet losses and error propagation can be jointly treated on the feature vectors extracted by the speech recognizer. In particular, we proposed solutions based on a Bayesian MMSE (Minimum Mean Square Error) estimation of those feature vectors affected by packet losses. In general, this estimate can be expressed as $\hat{\mathbf{x}}_t = E[\boldsymbol{x}_t|\Lambda]$, where $\boldsymbol{x}_t$ corresponds to the original feature vector and Λ represents the *a priori* knowledge about $\boldsymbol{x}_t$ [8, 9, 10]. Thus, the corrupted feature vectors are replaced by the expected value of the uncorrupted ones given some additional information Λ. This information consists of those feature vectors before and after a given loss, which is refined using a hidden Markov model (HMM). This model allows us to consider the distortions introduced by packet losses (including error propagation) and the temporal correlations of the speech signal [11, 12, 13].

One of the virtues of our MMSE estimate is that we can obtain some information on the confidence associated with each of the reconstructions performed. Additionally, this information can be used by recognizer-based techniques, such as the Soft-Data approach and the Weighted Viterbi algorithm (WVA), in order to consider this uncertainty in the recognition process [14]. These techniques require that the recognizer must be fed, as usual, with the feature vectors provided by the PLC block plus a reliability measure for those features. In this Ph.D. dissertation we proposed different techniques to compute the reliability factors, carrying out a comparative study of the results obtained by both approaches [11].

The architecture NSR can be modified in order to extract the feature vectors from the codec parameters (see Fig. 1a). This variant, called B-NSR (Bitstream-based NSR) or transparameterization, avoids the speech signal reconstruction by introducing a bitstream-based feature extraction that directly transforms the received codec parameters into recognition features. There are several reasons why the B-NSR approach can be attractive, among which are the following:

- Speech codecs usually include some type of post-processing at the decoder in order to obtain a decoded signal perceptually improved. However, this post-processing is not optimized for an objective performance measure as in speech recognition.
- It is not necessary to reconstruct the speech signal. This provides a computational saving.

For these reasons, in this Ph.D. thesis we proposed several transparameterization schemes for different popular speech codecs [11, 15]. This approach also allows us to develop an efficient adaptation, in terms of word accuracy and computational resources, of those techniques based on MMSE estimation and uncertainty treatment (soft-data and WVA), which were described above [11].

3. Experimental results

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group in [16]. The Aurora DSR front-end [17] provides a 14-dimension feature vector containing 13 MFCC (Mel Frequency Cepstral Coefficients) plus log-Energy. Furthermore, these vectors are extended by appending the first and second derivatives of the features. The recognizer is the one provided by Aurora and uses 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 training and testing data are extracted from the Aurora-2 database (connected digits). Training is performed with 8400 clean sentences and test is carried out over set *A* (4004 clean sentences distributed into 4 subsets).

In this work we have used two widely used CELP-based codecs: G.729A and AMR (Adaptive Multi-Rate) mode 12.2 kbps. In addition, iLBC (internet Low Bit-rate Codec) is also included because its design is oriented to increase the robustness against packet losses.

The channel burstiness exhibited by lossy packet networks was modelled by a 2-state Markov model. In particular, we obtained a wide set of channel conditions from a packet loss rate of 0% (condition C0) to 20% with a mean burst duration of 4 consecutive packets (condition C4), which simulate realistic situations of wired and wireless channels.

Since this paper is an extended abstract of a wide research work, we will only show the speech recognition results obtained by the best of our proposals. However, further readers can find a complete analysis of results in [2]. Table 1 collects a brief summary of our experimental results. In first place, we show the baseline results for G.729, AMR 12.2 and iLBC. These results correspond to carry out the recognition task using directly the decoded speech. On the other hand, the following rows show the results obtained by the combination of a B-NSR scheme, as feature extraction method, and the best proposed receiver-based PLC technique. In particular, this PLC technique carries

out the recontruction of those feature vectors affected by packet losses by means of an MMSE estimation, which additionally assigns confidence values to the estimates. These confidence values are taken into account during the recognition process to achieve further improvements. The last row shows the results obtained by the DSR approach described in [17], which can be considered as an upper limit for NSR. In this comparison, we must highlight that our proposals clearly outperform the speech recognition from decoded speech and they reduce considerably the differences between NSR and DSR.

4. Conclusions

This research work presents a thorough study of the degradations suffered by coded-speech recognition systems in presence of packet losses. In this sense, we have identified a new degradation source that is given by the inter-frame dependencies introduced by speech codecs. Thus, when a packet loss happens, in addition to the loss of information, a propagated error appears after the loss. In order to tackle these problems, we have proposed a wide range of packet loss concealment techniques based on the sender and the receiver. The first group of techniques modifying the structure of the client in order to improve the perceptual quality of the speech synthesis and, subsequently, the speech recognition accuracy. On the contrary, the proposed receiver-based techniques focus exclusively on improving the speech recognition accuracy by replacing those features vectors affected by packet losses and considering the remaining uncertainty in the recognizer. This Ph.D. dissertation, in conclusion, just to facilitate the understanding of degradation suffered by speech codecs in presence of packet losses and the wide range of proposed solutions, should serve as a reference work to develop new packet loss concealment techniques.

5. Acknowledgments

This work was supported by the Spanish MEC in the project FEDER TEC2007-66600 and the national plan for mobility of researchers.

6. References

[1] A.M. Peinado and J.C. Segura. *Speech recognition over digital channels. Robustness and standards*. Wiley, 2006.

[2] J.L. Carmona. *Coded-speech recognition over IP networks*. Ph.D. thesis, University of Granada, 2009. `http://digibug.ugr.es/bitstream/10481/2349/1/18056052.pdf`

[3] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, V. Sánchez, A.M. Gómez, "Rendimiento perceptual y reconocimiento con codificadores VoIP sobre redes de paquetes," *Actas de las IV Jornadas de la Red de Tecnologas del Habla*, pp. 249-254, Zaragoza (Spain), 2006.

[4] A.M. Gómez, J.L. Carmona, A.M. Peinado, V. Sánchez, "Intelligibility evaluation of Ramsey-derived interleavers for Internet voice streaming with the iLBC codec," *in proceedings of Interspeech 2008*, Brisbane (Australia), 2008.

[5] J.L. Carmona, J.L. Pérez-Córdoba, A.M. Peinado, A.M. Gómez, J.A. González, "A scalable coding scheme based on interframe dependency limitation," *in proceedings of ICASSP'08*, pp. 4805-4809, Las Vegas (USA), 2008.

[6] A.M. Gómez, J.L. Carmona, A.M. Peinado, V. Sánchez, "A multipulse-based forward error correction technique for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 8, No. 6, pp. 1258-1268, 2010.

[7] J.L. Carmona, A.M. Gómez, A.M. Peinado, J.L. Pérez-Córdoba, J.A. Gonzaĺez, "A multipulse FEC scheme based on amplitude estimation for CELP codecs over packet networks,"*in proceedings of Interspeech 2010*, Makuhari (Japan), Septiembre, 2010.

[8] A.M. Gómez, A.M. Peinado, V. Sánchez, J.L. Carmona, "Interleaving and MMSE estimation with VQ replicas for distributed speech recognition over lossy packet networks," *in proceedings of Interspeech 2006*, Pittsburgh (USA), 2006.

[9] J.A. González, A.M. Peinado, A.M. Gómez, J.L. Carmona, J.A. Morales, "Efficient VQ-based MMSE estimation for robust speech recognition," *in proceedings of ICASSP'10*, Dallas (USA), 2010.

[10] J.A. González, A.M. Peinado, A.M. Gómez, J.L. Carmona, "Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, in press, pp. 1-15, 2010.

[11] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba and A.M. Gómez, "MMSE-based packet loss concealment for CELP-coded speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, Vol. 8, No. 6, pp. 1341-1353, 2010.

[12] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, A.M. Gómez, J.A. González,"Robust encoded speech recognition over IP networks,"*in proceedings of SIMPE 2010*, Lisbon (Portugal), Septiembre, 2010.

[13] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, A.M. Gómez, J.A. González,"Dealing with acoustic noise and packet loss in VoIP recognition systems," *to appear in proceedings of FALA 2010*, Vigo (Spain), 2010.

[14] A.M. Gómez, A.M. Peinado, V. Sánchez, J.L. Carmona, "A robust scheme for distributed speech recognition over loss-prone packet channels," *Speech Communication*, Vol. 51, pp. 390-400, 2009.

[15] J.L. Carmona, A.M. Peinado, J.L. Pérez-Córdoba, V. Sánchez and A.M. Gómez, "iLBC-based transparameterization: A real alternative to DSR for speech recognition over packet networks," *in proceedings of ICASSP'07*, Honolulu, USA, 2007.

[16] H.G. Hirsh and D. Pearce. *The Aurora experimental framework for the performance evaluations of speech recognitions systems under noise conditions*. ISCA ITRW ASR, 2000.

[17] ETSI ES 201 108. *Front-end feature extraction algorithm; Compression algorithms*, 2000.