Dealing with Acoustic Noise and Packet Loss in VoIP Recognition Systems

José L. Carmona, Antonio M. Peinado, José L. Pérez-Córdoba, José A. González, Angel M. Gómez

Dpto. Teoría de la Señal, Telemática y Comunicaciones, University of Granada

{maqueda, amp, jlpc, joseangl, amgg} @ ugr.es

Abstract

In this paper the robustness of Network Speech Recognition (NSR) systems is analyzed. In NSR the speech signal is transmitted using a conventional speech codec from the client to the server, where the recognition task is carried out. The use of speech codecs degrades the performance of such systems, mainly in presence of acoustic noise and packet losses. First, we study the effects of possible degradation sources. Then, we propose a new NSR solution based on a robust feature extractor and an efficient packet loss concealment (PLC) algorithm, which compensates the possible degradations by means of a cepstral compensation and linear interpolation. The experimental results are obtained for a well-known speech codec, AMR 12.2 kbps, using a noisy database (Aurora-2) and several packet loss conditions. The results show that our proposal achieves noticeable improvements over the baseline results.

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

1. Introduction

IP Packet switching networks have originated a global network of networks (Internet). Voice transmission over this type of network, called Voice over IP (VoIP), has shown strong growth during the past years, and it has turned into one of the key aspects of the current state of telecommunications. In parallel with voice and data convergence provided by VoIP platforms, new standards of wireless Internet access have led to a convergence of IP and mobile telephony networks. This paradigm will give rise to a new concept of nomadic access, hybrid of fixed and mobile access, linked to the incorporation of IP technologies and provided by suppliers of these new technologies.

Under this parardigm, automatic speech recognition offers a natural oral interaction and fast access to information. Unfortunately, there are several problems to implement a powerful automatic speech recognition subsystem into mobile terminals due to their size restrictions and limited computation capacity. Distributed speech recognition (DSR) avoids these hardware constraints by placing the most complex computational requirements of speech recognition into a remote server [1]. Moreover, the structure of a remote recognition system is well suited for the IP model, since it is the provider who implements the recognizer depending on its needs.

Although during the last years several DSR standards have been issued [2, 3], the lack of DSR codecs in the existing devices supposes a barrier for its deployment. Thus, most of the current DSR systems employ a conventional speech codec in order to transmit the speech signal to the server, where the recognition task is performed. This architecture is also known as network-based speech recognition (NSR), since the whole speech recognizer resides in the network from the client's point of view. NSR does not require any modification in the client terminal, since it uses deployed VoIP platforms. However, speech coding involves an information loss that may reduce speech recognition performance. Moreover, we have to take into account this performance reduction in presence of other implicit problems of remote speech recognition, such as acoustic noise (the acoustic context of the terminal may vary) and degradations introduced by the communication channel (packet loss for IP networks) [1].

This paper focuses on analyzing the impact of acoustic noise and packet losses on NSR systems. The NSR architecture based on decoded speech allows us to employ robust feature extractors in adverse acoustic conditions, such as the advanced front-end (AFE) proposed by the Aurora-2 working group [3]. However, as we will see, packet loss involves a drastic performance reduction in this kind of NSR systems. Thus, an analysis of the possible degradation sources is carried out prior to propose new solutions to the packet loss problem for NSR architecture working in noisy acoustic conditions.

The structure of this paper is the following. In Section II we present the experimental framework. Section III is devoted to analyze the possible degradation sources in robust NSR systems. In Section IV we propose a new framework in order to increase the robustness of NSR systems in adverse acoustic and channel conditions. Finally, in Section V we summarize our conclusions.

2. Experimental framework

The experimental setup is based on the framework proposed by the ETSI STQ-Aurora working group for the Aurora-2 database [4]. The Advanced front-end [3] 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. This test set contains 4 subsets (1001 sentences each) contaminated with four different types of additive noise (subway, babble, car and exhibition) at different SNRs (clean, 20, 15, 10, 5, 0 and -5 dB). For every SNR, the word accuracy (WAcc) is obtained by averaging the word accuracies of the four subsets. A mean word accuracy is computed by averaging the results obtained for all the SNRs excluding those of clean and -5 dB.

In the analysis of possible degradation sources carried out in this paper, we have used two widely used CELP-based codecs: G.729A [5] and AMR (Adaptive Multi-Rate) [6]. In addition, iLBC (internet Low Bit-rate Codec) [7] is also included, since its design is oriented to increase the robustness against packet losses.

The channel burstiness exhibited by lossy packet networks

Condition	p	q	L_{avr}	PL_r
C0	0	–	–	0%
C1	0.0526	1.0000	1	5%
C2	0.0555	0.5000	2	10%
C3	0.0588	0.3333	3	15%
C4	0.0625	0.2500	4	20%

Table 1: Packet loss conditions.

is modeled by a 2-state Markov model. The transition probabilities between states, p and q, can be set according to an average burst length (L_{avr}) and a packet loss ratio (PL_r). The performance of the NSR systems presented in this paper is tested under the channel conditions listed in Table 1.

3. Effect of Channel and Acoustic Degradations on NSR systems

The performance of NSR systems will be determined by the intrinsic codec robustness. Table 2 shows word accuracy (WAcc) results obtained by different remote speech recognition systems in packet loss conditions. The results are obtained using two training stages. The first one, called T1, refers to train the speech recognizer using original speech, i.e. non-coded speech. The second one, labeled T2, corresponds to carry out the training stage with decoded speech. As shown, the results are consistently higher for training T2, since using decoded speech in training reduces the mismatch in testing. The results obtained by the DSR system defined in [3] are included as reference. This system obtains noticeable improvements respect to NSR systems since it is specifically oriented to remote speech recognition. The WAcc result obtained directly from original speech, i.e. without using any coding scheme, is 87.74 and, thereby, the quantization stage used in [3] does not involve any performance reduction. On the contrary, there exists some performance reduction for NSR systems, although it is somewhat alleviated when the training is carried with decoded speech (T2). The speech codecs based on CELP do not achieve an optimum performance in lossy channel conditions because they use predictive techniques. For example, G.729 achieves good results in clean channel conditions transmitting linear spectrum pair (LSP) coefficients by means of a differential predictive quantifier. However, this strategy makes the codec more vulnerable to consecutive packet losses, since once a packet loss is finished, the LSP prediction is still significantly degraded. This justifies that even AMR (4.75 kbps) achieves better results than G.729 (8 kbps) for non-ideal conditions. iLBC tackles these problems by removing all types of inter-frame dependencies in the encoding process [8]. However, the price to pay is a considerable increase of bit-rate (15.2 kbps). In general, the performance of NSR is particularly lower than that of DSR when packet losses are grouped in bursts.

Speech decoders try to reduce the perceptual impact of packet losses by means of packet loss concealment (PLC) algorithms. These algorithms are usually based on repetition and progressive muting of the last received speech segment. The purpose of repetition is to conceal the effect of lost frames, whilst the progressive muting avoids the generation of annoying sounds in case of several consecutive lost frames. Nevertheless, this progressive muting leads to an increase on the insertion errors in the recognizer (artificial silences). In addition, we can observe a degradation of the decoded speech signal corresponding to correct frames after a packet loss. This degradation is inherent to the predictive nature of the encoding process of CELP-based codecs, such as G.729 and AMR 12.2 kbps. Moreover, when the decoded speech signal is used as input of a robust feature extractor, such as AFE, the error propagation is strengthened. In particular, AFE includes a noise reduction block which estimates noise characteristics in order to reduce its negative effect. In this sense, packet losses prevent this estimation process, hence a new source of degradation appears.

Our objective now is to distinguish the effects of the 'repetition and muting' effect during the burst from the 'codec error propagation' and 'AFE error propagation' after the burst. In order to do so, we can study the impact of each type of degradation on the reduction of the recognition accuracy. The following experiments are carried out by substituting speech samples and feature vectors by those corresponding to a clean transmission using AMR 12.2 kbps.

1. 'Repetition and Muting' experiment: Speech samples belonging to correctly received frames are replaced by the corresponding correct samples. Thus, the only degraded samples that remain are those where the repetition and muting algorithm is applied.
2. 'Codec Error Propagation' experiment: Speech samples belonging to lost frames are replaced by their corresponding correct samples, so that the error propagated by the speech decoder is the only remaining degradation.
3. 'AFE Error Propagation' experiment: In this case, a double substitution is carried out. First, we follow the same procedure as in the 'Repetition and Muting' experiment (that is, to replace those samples affected by codec error propagation). Second, the extracted feature vectors corresponding to the lost packets are replaced by the original ones. Thus, the remaining degradation is mainly due to the corruption of the internal states of AFE after a packet loss.

The results of these experiments are presented in Table 3. As shown, the main source of degradation is given by the 'repetition and muting' effect. PLC algorithms included in speech decoders are based on perceptual considerations that are unsuitable for recognition tasks. Nonetheless, both propagation effects that appear after a packet loss are also a considerable source of degradation. As can be observed, codec error propagation reduces the speech recognition performance at all SNR conditions, whilst AFE error propagation is more appreciable for low SNR conditions. These results are consistent, since good noise estimations are more significant for those test conditions with low SNRs. Note that AFE error propagation does not reduce the recognition performance in clean acoustic conditions, while its effect is more detrimental than codec error propagation for those conditions with SNR below 5 dB.

4. Improving NSR from decoded speech

In this section we propose a new scheme that allows us to reduce the impact of the degradation sources studied in the previous section. First, we will introduce some modifications in the Advanced Front-End (AFE) in order to make their spectral estimates more robust against packet losses. Later, we will describe a packet loss concealment oriented to speech recognition, and a compensation technique that reduces the impact of the remaining error propagation. We will test the performance of our proposal using AMR 12.2 kbps.

		iLBC		*G.729*		*AMR 12.2*		*AMR 7.95*		*AMR 4.75*		DSR	
Training		*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*	*T1*	*T2*
Conditions	*C0*	84.78	86.80	84.22	85.23	85.47	86.07	83.75	83.96	82.06	83.48	87.39	87.81
	C1	83.89	86.12	76.91	78.31	81.62	82.92	79.63	80.56	77.78	79.82	87.25	87.61
	C2	80.70	82.89	68.30	68.75	76.49	77.56	74.50	74.88	72.68	74.30	86.04	86.47
	C3	76.14	78.36	62.98	63.07	70.89	71.81	69.08	68.90	67.78	68.36	83.69	84.13
	C4	71.41	73.51	58.48	58.40	65.54	65.75	63.85	62.83	62.74	62.51	79.79	80.35

Table 2: Recognition accuracy (WAcc (%)) for NSR systems based on different speech codecs.

	Chan.	SNR							
	Cond.	*Clean*	*20 dB*	*15 dB*	*10 dB*	*5 dB*	*0 dB*	*-5 dB*	*Avg.*
AMR 12.2	C0	99.13	97.90	96.61	92.38	82.98	60.48	28.90	86.07
	C4	84.10	81.11	78.00	71.50	59.06	39.07	17.73	65.75
Repetition and Muting	C4	85.95	83.69	80.79	74.95	62.10	40.29	16.41	68.36
Codec Error Propagation	C4	97.69	95.49	93.54	89.32	78.96	56.79	26.42	82.82
AFE Error Propagation	C4	99.13	97.47	95.57	90.22	77.99	51.82	20.71	82.62

Table 3: Recognition accuracy (WAcc (%)) for substitution experiments using AMR 12.2 kbps and test A of Aurora-2 (clean training).

4.1. Modified AFE

In comparison with non-advanced front-ends, the AFE standard introduces two noise reduction techniques based on a Wiener filtering (WF) and an SNR-dependent waveform processing. Since the WF block is the main noise reduction technique applied in this ETSI standard we will focus on it. In particular, AFE is based on a two-stage mel-warped WF technique [9]. Its basic principle is a double WF filtering (the output of the first stage is the input to the second one). The WF filter is computed for every block of $M = 80$ samples. In order to obtain the WF coefficients is necessary to buffer 4 blocks of samples and to compute an estimate of the power spectral density (PSD) using frames of length $N_{in} = 200$ samples (between the samples 60 and 259 of the input buffer). Obviously, the buffer necessary for the WF design produces an expansion of the impact of packet loss as shown in Figure 1. The speech codec frame division is represented in the upper part. A 20-ms frame duration has been assumed, such as AMR 12.2 kbps. The two WF stages included in the noise reduction block work on subframes of 80 samples, which are represented in the second line. Finally, the bottom of the diagram represents the way that feature vectors (FV frames) are extracted from the denoised samples. Feature vectors are computed from overlapping speech segments of 25-ms length (200 samples) and 10-ms frame shift. In addition the diagram shows how subframes and feature vectors are affected by one packet loss. As can be seen, a packet loss corresponding to a speech frame of 20 ms can affect upto 9 feature vectors.

In practice the PSD estimation is carried out after applying a Hanning window of 200 samples over those samples stored in each buffer. We can assume that spectral estimates are mainly obtained using only the two central subframes of each buffer. For this reason, we can consider that a loss actually affects 7 feature vectors and its effects are more detrimental when the second stage of the noise reduction block is involved. In order to reduce the corruption of the spectral estimates, we define two non-updating indicators (NUI) by means of the following mapping functions,

$$\mathrm{NUI}_1(n) = \begin{cases} 1 & \text{if } \mathrm{PLI}(m) = 1, m \in \left(\lceil \frac{n+1}{2} \rceil, \lceil \frac{n+4}{2} \rceil\right) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\mathrm{NUI}_2(n) = \begin{cases} 1 & \text{if } \mathrm{PLI}(m) = 1, m \in \left(\lceil \frac{n-1}{2} \rceil, \lceil \frac{n+2}{2} \rceil\right) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

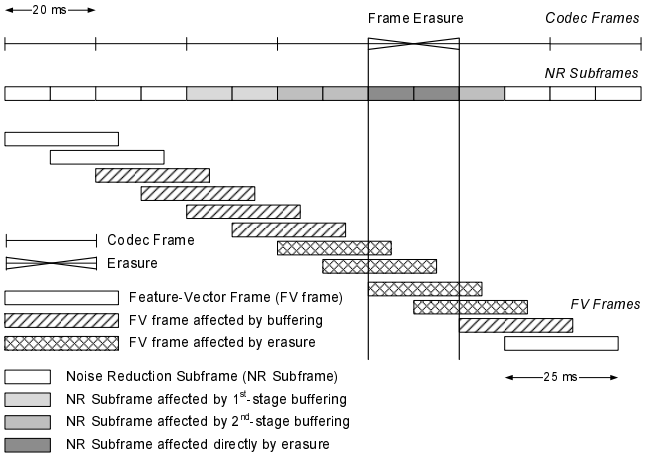

Figure 1: Feature vectors affected by a frame erasure.

where $\mathrm{PLI}(m)$ is the packet loss indicator for the codec frame m (1 for packet loss, 0 otherwise), and n is the time index of a given noise-reduction subframe. $\mathrm{NUI}_i(n) = 1$ indicates that the PSD estimate of the ith stage must not be updated for the subframe with index n, while $\mathrm{NUI}_i(n) = 0$ corresponds to normal updating.

4.2. PLC Algorithm

Although AFE propagation error is limited thanks to the modifications proposed in the previous subsection, there will still be a remaining degradation because spectral estimates have not been updated. In addition, we also have to consider that the decoded speech signal will be corrupted after a burst due to the codec error propagation [10]. We can assume that this degradation behaves as an additive noise that affects those frames after a burst. Under this approach, we can compensate this degradation by means of a cepstral normalization. In particular, we have shown [11, 10] that this remaining degradation can be effectively compensated by means of FCDCN (Fixed Codeword-Dependent Cepstral Normalization). The principle of this technique is to apply an additive correction vector $\mathbf{r}$ to the noisy feature vector $\mathbf{y}$ that depends on the length of the burst, the po-

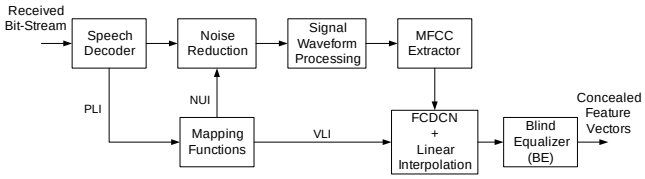

Figure 2: Block diagram of the proposed robust feature extractor including PLC techniques.

	Channel Conditions				
PLC Algorithms	*C0*	*C1*	*C2*	*C3*	*C4*
AFE	86.07	82.92	77.56	71.81	65.75
AFE+LI	–	82.68	78.31	72.87	66.99
MAFE	–	82.98	78.12	73.29	68.13
MAFE+LI	–	83.1	79.5	75.36	70.69
MAFE+FCDCN+LI	–	83.89	80.48	76.18	71.36

Table 4: WAcc results for different PLC techniques using AMR 12.2 kbps over test A of Aurora-2 (clean training).

sition after the burst, and vector **y** itself. Since the correction depends on the observed vector, this is quantized and a compensation is computed for every quantizer cell during a stereo training. This estimation can be carried out for different burst lengths by simulating as many frame erasures as needed in order to obtain a accurate compensation. Further details about this PLC compensation can be found in [11]. In order to obtain a fine representation of the cepstral space, we have used split vector quantizers with the same number of centroids than those ones employed in the ETSI DSR standard [3]. As in [11], we have considered 20 positions after every loss and a maximum burst length of 5 frames.

In addition to FCDCN compensation, it is necessary to define a PLC algorithm in order to substitute those feature vectors affected directly by a frame erasure (see Figure 1).Thus, feature vectors corresponding to lost frames can be reconstructed by means of a simple linear interpolation between the last and first correct vectors before and after a packet loss,

$$\hat{\mathbf{x}}(t) = \mathbf{x}(t_s) + \frac{\mathbf{x}(t_e) - \mathbf{x}(t_s)}{t_e - t_s}(t - t_s) \qquad t_s < t < t_e \quad (3)$$

where $\hat{\mathbf{x}}(t)$ is the estimated feature vector at time t, and $\mathbf{x}(t_s)$ and $\mathbf{x}(t_e)$ are the last and first correct feature vectors before and after a burst, respectively. This technique has proven to be more powerful than the repetition of the nearest neighbor vector in NSR systems [10].

Figure 2 presents a block diagram of the proposed feature extractor, where the feature vectors corresponding directly to a loss are marked by a vector loss indicator (VLI). As shown, PLC techniques are inserted between the AFE blocks (noise reduction, signal waveform processing, MFCC extractor and blind equalizer). Table 4 shows the results obtained with AMR 12.2 kbps using the proposed PLC techniques. The baseline (AFE) corresponds to carrying out the recognition task from decoded speech including the PLC algorithm defined by the legacy codec. The second row shows the results obtained by applying only linear interpolation (AFE+LI). The algorithms named MAFE refer to those solutions based on the modified AFE explained in the previous section, which uses non-updating indicators in order to avoid the corruption of the AFE internal states. Finally, the results labeled as MAFE+FCDCN+LI correspond to carrying out the FCDCN compensation and linear interpolation described in this section. As shown, this last approach achieves the best results.

5. Conclusions

In this work we have analyzed the robustness of NSR systems in adverse acoustic and transmission channel conditions. In particular the analyzed NSR architecture is based on the use of decoded speech as input of the advanced front-end (AFE) defined by ETSI. First, we have identified three possible sources of degradation when a packet loss appears. These can be summarized as follows. The first one is generated by the PLC algorithms included in the speech decoder. These PLC algorithms are usually based on perceptual considerations that are not appropriate for speech recognition. The second one is the error propagation associated to those speech codecs based on the CELP paradigm. The third source of degradation is caused by the corruption of the internal states (spectral estimates) of AFE during a loss burst. Secondly, we have proposed a new framework in order to reduce the impact of these degradation sources. Our proposal is based on a modified version of AFE, which partially avoids the corruption of its internal states, and a PLC algorithm oriented to speech recognition, which is based on a cepstral compensation technique and linear interpolation.

6. Acknowledgments

This work was supported by the Spanish MEC in the project FEDER TEC2007-66600.

7. References

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

[2] ETSI ES 201 108. “Front-end feature extraction algorithm; Compression algorithms”, 2000.

[3] ETSI ES 202 050. “Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms”, 2005.

[4] 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.

[5] Recommendation ITU-T G.729. “Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited-linear-prediction”, 1996.

[6] 3GPP TS 26090. “AMR speech codec; Transcoding functions”, 1999.

[7] S. Andersen, W. Kleijn, R. Hagen, J. Linden, M. Murthi, and J. Skoglund, “iLBC - a linear predictive coder with robustness to packet losses”. in IEEE Workshop on Speech Coding, 2002.

[8] 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 Proc. of ICASSP’07, Honolulu, USA, 2007.

[9] A. Argawal and Y. Cheng. “Two-stage mel-warped Wiener filter for robust speech recognition”, in Proc. of ASRU’99, Keystone, USA, 2009.

[10] 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., in Press. Online: http://dx.doi.org/10.1109/TASL.2009.2033891, accessed on 18 Jun 2010.

[11] A.M. Gómez, A.M. Peinado, V. Sánchez and A.J. Rubio, “Recognition of coded speech transmitted over wireless channels”, IEEE Trans. Wireless Commun., Vol. 5, No. 9, 2006.