# iLBC-BASED TRANSPARAMETRIZATION: A REAL ALTERNATIVE TO DSR FOR SPEECH RECOGNITION OVER PACKET NETWORKS

*José L. Carmona, Antonio M. Peinado, José L. Pérez-Córdoba, Angel M. Gómez, Victoria Sánchez*

Dpt. Teoría de la Señal, Telemática y Comunicaciones, University of Granada
maqueda@ugr.es

**ABSTRACT**

This paper proposes a method for the remote recognition of speech coded with the iLBC codec, which is employed by a number of VoIP systems. While the usual way of performing recognition of coded speech is to decode first the speech signal and use it as input to the recognition engine, our system directly converts the iLBC parameters into recognition features. The main advantage of this approach is to avoid any type of decoding post-processing which, although originally conceived to improve the speech perception, can be harmful for a recognition system. Our method ensures the compatibility between the speech spectra provided by the iLBC codec and those employed for cepstrum computation and introduces a robust and suitable packet loss concealment strategy. Our experimental results show that the proposed system achieves a performance better than that obtained from iLBC-decoded speech and similar to that of a distributed speech recognition system over a clean or degraded transmission channel.

***Index Terms***— Speech recognition, speech codecs, iLBC codec, packet network, NSR.

## 1. INTRODUCTION

The deployment of new mobile services accessed by smart interfaces offers new chances to speech technologies, which can allow a natural, ubiquitous and pervasive multimodal interaction. In particular, automatic speech recognition (ASR) plays a preeminent role among these technologies since it is the basic element for human-machine communication. Although ASR systems can be embedded in portable devices, a client/server architecture, where the speech signal (or some speech parameters) is encoded by the local client while the final recognition is carried out in a remote server, has several advantages. First, it involves a simpler client. Also, it frees the user of upgrading and maintenance tasks, and allows easy language portability.

There currently exist two different architectures for remote speech recognition (shown in figure 1) which basically differ in the way that speech encoding (and decoding) is carried out [1]:

- **Network speech recognition** (NSR): the speech signal is compressed by means of a speech codec.
- **Distributed speech recognition** (DSR): the speech signal is processed by a local front-end which directly obtains the specific features (usually cepstrum plus energy) used by the remote server to perform recognition, avoiding so the speech coding/decoding process required by NSR.

Work supported by projects MEC/FEDER TEC2004-03829/TCM and FIT-330503-2006-2.

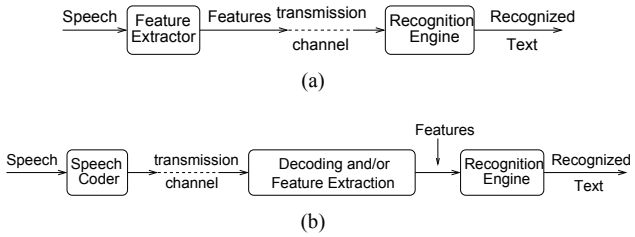


**Fig. 1**. Block diagrams of (a) a distributed speech recognition (DSR) system and (b) a network-based speech recognition (NSR) system.

During the last years, DSR has attracted the attention of a number of researchers since it offers several advantages such as a smaller bit-rate or increased robustness against transmission errors [2]. However, NSR also has some interesting characteristics. For example, it does not require the introduction of new codecs, so that it is possible to employ the same as those employed in mobile phones or VoIP (Voice over IP) terminals, depending on the application. Also, it allows a natural, high-quality reconstruction of the original speech, which can be useful in certain applications where speaker identification may be required.

In this paper we deal with the development of a NSR system which can compete, in terms of accuracy and robustness against transmission errors, with DSR. Due to the convergence of networks towards IP, we will consider a packet channel where packet loss is the main source of degradation. There is a number of speech codecs oriented to VoIP such as G.723.1, G.729 or, more recently, iLBC [3]. We have selected iLBC since it does not introduce inter-frame dependencies, so that it is more robust against packet losses and provides us with a good starting point. iLBC coder has two operation modes: 20 ms frame mode (15.2 kbps) and 30 ms frame mode (13.33 kbps). Our proposal is developed over the first mode.

Although the most straightforward way for NSR is to perform recognition from the decoded signal, it is also possible to extract the recognition features directly from the codec parameters without reconstructing the speech. This approach [5, 6], usually referred to as *transparametrization* or *Bitstream-based NSR* (B-NSR), has several advantages with respect to basic NSR. First, since there is no speech signal reconstruction, B-NSR does not suffer from the inter-frame dependencies typical of LPC-based codecs. Furthermore, it avoids the artifacts introduced during decoding to improve the perceptual quality for a better speech perception or to mitigate annoying effects. In particular, it avoids the 'substitution and muting' packet loss concealment (PLC) mechanism (typically applied in the case of lost frames due to channel degradation)

since it is quite harmful in an ASR application [6, 7]. In its place, a PLC algorithm suitable for speech recognition may be employed. Thereby, the robustness of the whole system against channel degradation can be considerably improved.

In the next sections, we propose and develop a transparametrization method from the iLBC parameters to those employed in DSR. As it is shown, especial care must be taken to ensure the compatibility of the iLBC and DSR spectra. Also, a PLC scheme which employs repetition and interpolation is introduced.

## 2. iLBC TRANSPARAMETRIZATION APPROACH

### 2.1. Aurora framework

The proposed transparametrization converts iLBC codec parameters into feature vectors as defined in the Aurora ETSI standards for DSR. They consist of 13 MFCC coefficients plus the log-energy. Dynamic features are computed at the recognition stage.

The recognizer is the one provided by Aurora [4] and 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 data base (connected digits). The training is performed with 8440 clean sentences and the test is carried out over the *set A* (4004 clean sentences distributed into 4 subsets). The vocabulary is made up of 11 digits between 0 and 9 (zero has two sound descriptions: 'zero' and 'o'). The mean length of each sentence is 1.5 s. The recognition measure used was Word Accuracy ($WAcc(\%)$).

### 2.2. Operative principle

The iLBC codec (mode 15.2 kbps) operates on speech frames of 160 samples which are divided into four sub-frames. Each iLBC frame contains one set of LSFs (Line Spectrum Frequencies) obtained from a 10th order linear prediction analysis carried out once every frame using an asymmetric window centered in the third subframe. On the other hand, the DSR feature extraction algorithm is performed over 200 samples (25 ms) every 80 samples. Because of the differences in the speech signal analysis between DSR and iLBC, the following interpolation of the LSF coefficients is applied.

$$\begin{aligned} \overline{LSF}_{2\cdot n} &= \frac{6\cdot LSF_{n-1}+13\cdot LSF_n+1\cdot LSF_{n+1}}{20} \\ \overline{LSF}_{2\cdot n+1} &= \frac{1\cdot LSF_{n-1}+13\cdot LSF_n+6\cdot LSF_{n+1}}{20} \end{aligned} \qquad n=1,2,\ldots \quad (1)$$

where $\overline{LSF}_{2\cdot n}$ and $\overline{LSF}_{2\cdot n+1}$ are the LSF sets of the DSR frames $2n$ y $2n+1$ and $LSF_n$ is the LSF set of the iLBC frame $n$. Thereby, we double the number of $LSF$ sets provided by the iLBC coder.

The MFCC coefficients can be computed following the DSR standard replacing the FFT spectrum by the following LPC spectrum,

$$|H'(\omega_i)| = \sigma|H(\omega_i)| = \frac{\sigma}{1+\sum_{i=1}^{10} a(k)e^{-j\omega_i k}} \qquad (2)$$

where $\sigma$ is the LPC gain, $\omega_i = 2\pi i/N$ $(i = 0, ..., N-1)$ and $|H(\omega_i)|$ is the gain-normalized LPC spectrum evaluated with $N = 256$ points [7].

At this point, a remarkable characteristic of the LPC analysis in the iLBC coder has to be discussed because of its effect in the proposed approach. That is the spectrum expansion, as shown in figure 2, which is performed according to,

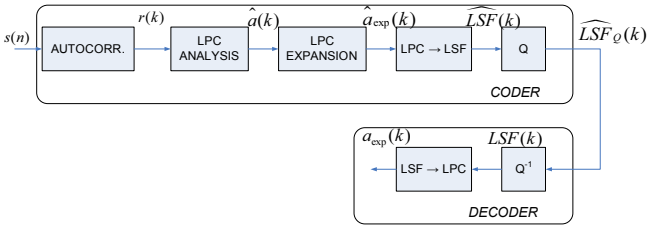


**Fig. 2**. Coding and decoding of the LSF coefficients in the iLBC codec.

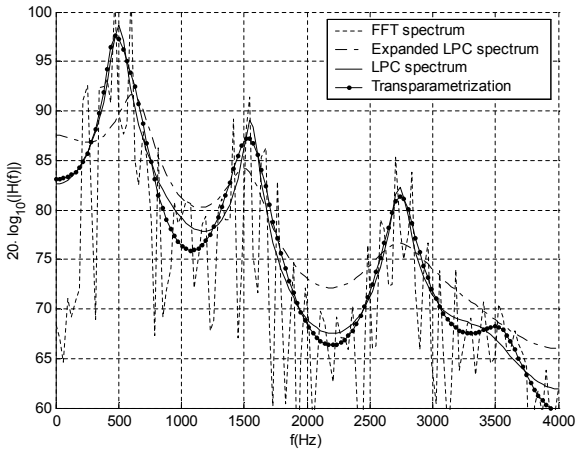


**Fig. 3**. Distortion produced by the spectrum expansion in the LPC analysis and the proposed approximation used by the transparametrization approach.

$$H_{exp}(z) = H(\frac{z}{\gamma}) \textbf{ or } a_{exp}(k) = a(k)\gamma^k \quad k = 0, ..., 10 \qquad (3)$$

where $\gamma$ is the expansion factor (equal to 0.9) and $H_{exp}(z)$ is the expanded LPC spectrum. This operation has several effects (see figure 3). The dynamic range of the spectrum is reduced, therefore the length of the impulse response of the synthesis filter is reduced too. Therefore, in case of a packet loss, the filter does not excessively propagate the error in the filter memory. Furthermore, the location of the poles around the origin is compressed reducing the quantization space and ensuring the stability of the synthesis filter.

However, this expansion is a serious inconvenient for our proposal because it introduces a considerable distortion in the LPC spectrum. It can be argued that this expansion can be reversed in the decoder. Nevertheless, the LSF quantization process prevents this possibility in the decoder since it would lead to unstable LPC filters.

To cope with this situation it is needed to consider the spectral characteristic of the coded residual signal. In this way, the decoded residual signal is processed to obtain a new set of LPC parameters ($a_{res}(k)$) which characterizes the LPC spectrum of the residual signal $H_{res}(\omega_i)$. Now, we can use these coefficients to obtain an improved LPC spectrum estimation by means of the following expression,

$$|\hat{H}(\omega_i)| = H_{exp}(\omega_i) \cdot H_{res}(\omega_i) = \\ = \frac{1}{\left|1 + \sum_{l=1}^{10} a_{exp}(l) e^{-j\omega_i l}\right|} \cdot \frac{1}{\left|1 + \sum_{k=1}^{10} a_{res}(k) e^{-j\omega_i k}\right|} \quad (4)$$

This expression accounts for the non-flat shape of the residual spectrum envelope after the expansion operation. Figure 3 shows the new spectrum estimation $|\hat{H}(\omega_i)|$, with its corresponding gain $\hat{\sigma}$, marked as "Transparametrization". It is observed that approximates the FFT spectrum even better than the original LPC spectrum.

From this estimation and following the procedure given in [7], the MFCC coefficients can be obtained as:

$$MFCC(k) = \begin{cases} M \log \hat{\sigma} + MFCC'(k) & k = 0 \\ MFCC'(k) & k = 1, ..., 12 \end{cases} \quad (5)$$

where $MFCC'(k)$ are the cepstral coefficients corresponding to $|\hat{H}(\omega_i)|$ and $\hat{\sigma}$ is its gain.

At the decoder side, the energy $E_{res}$ and the corresponding LPC coefficients ($a_{res}(k)$) are computed from the residual signal at the corresponding DSR rate (every 10 ms in frames of 25 ms long). The gain $\hat{\sigma}$ is computed using the following expression:

$$\hat{\sigma}^2 = \frac{E_{res}}{\frac{1}{2\pi}\int_{-\pi}^{\pi} \left| \frac{1}{1 + \sum_{k=1}^{10} a_{res}(k) e^{-j\omega k}} \right|^2 d\omega} \quad (6)$$

Finally, the MFCC coefficients are calculated according (1) using the decoded LSF parameters and $\log E$, which is obtained with the following expression:

$$\log E = \log \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} |\hat{\sigma} \hat{H}'(\omega)|^2 d\omega \right) \quad (7)$$

The integration operations, which appear in expressions (6) and (7), are computed in the time domain by means of sums.

### 2.3. Mitigation of errors

In practice, all kind of errors due to an IP channel can be considered as packet losses. To alleviate these packet losses, a PLC algorithm is required. In the DSR Aurora standard a repetition scheme of the correct MFCC coefficients is implemented, since it achieves better results than interpolation.

In our proposal, a mitigation technique is implemented in the parameter domain using: $LSF_{res}$ (calculated from $a_{res}(k)$), $E_{res}$ and $\overline{LSF}$. Several combinations of mitigation techniques have been tested and we found that the highest performance was obtained using linear interpolation for parameters derived from the residual signal ($LSF_{res}$, $E_{res}$) and a repetition technique (similar to the DSR concealment algorithm) for $\overline{LSF}$.

An additional problem consists in deciding which parameter set must be discarded when an iLBC packet loss occurs. This problem is a consequence of the differences between frame lengths in the iLBC codec and the DSR standard. The best solution for this work is shown in figure 4. The objective is to keep the quasi-periodicity of the residual signal (in case of a voiced segment) in order to introduce the minimal distortion over the parameters derived from it ($LSF_{res}$ and $E_{res}$).

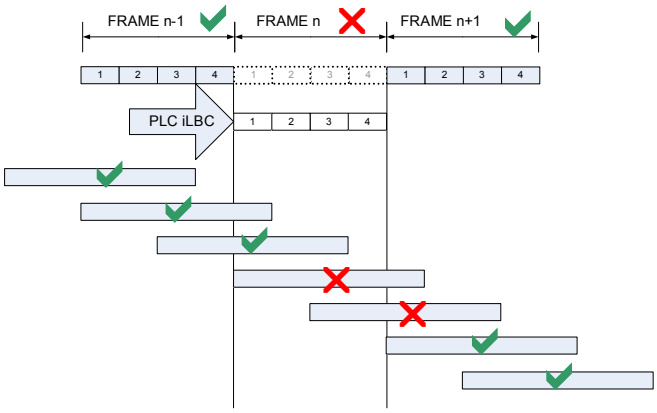


**Fig. 4**. Correspondence between iLBC lost packets (represented at the upper part of the figure) and interpolated parameters to be discarded at DSR rate (at the bottom part of the figure).

## 3. RESULTS

In this section performance results, in terms of word accuracy (*WAcc*), are reported. In order to emulate the behaviour of an IP channel, a two-state model (shown in figure 5) is used. This model is characterized by the transition probabilities $p$ and $q$ between the state 0 (packet received correctly) and the state 1 (packet lost). From these probabilities is easy to derive that,

$$P_{loss} = \frac{p}{p+q} \\ D_{burst} = \frac{1}{q} \quad (8)$$

where $P_{loss}$ is the loss rate and $D_{burst}$ is the mean loss burst length. The simulated channel conditions are obtained by selecting $P_{loss}(\%)$ and $D_{burst}$ from the sets [5 10 20 30 40 50] and [1 2 4 8 12 16], respectively.

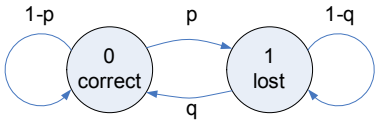


**Fig. 5**. Model of the IP channel.

As a reference, table 1 shows the *WAcc* results obtained from the synthesized speech using iLBC. The best recognition rate will be the *WAcc* value of the DSR Aurora system on clean channel conditions which marks an upper limit of 99.04%, while the results in clean conditions are 98.92% and 98.90% for iLBC NSR and iLBC B-NSR, respectively.

Table 2 shows recognition rates for the iLBC B-NSR system with the mitigation technique described in the previous section. Comparing the recognition rate values in both tables, it is clear that the transparametrization method outperforms the recognition system from synthesized speech for all channel conditions.

A comparison of the proposed iLBC-based B-NSR system with other NSR systems (using decoded speech), based on AMR 12.2 kbps and G.729 8 kbps using 1 and 2 frames per packet respectively, and DSR according to the ETSI standard, packing 2 frames per

| Loss Rate | Burst Length | | | | | |
|---|---|---|---|---|---|---|
| | *1* | *2* | *4* | *8* | *12* | *16* |
| *5%* | 98.34 | 97.45 | 96.10 | 94.65 | 94.24 | 93.88 |
| *10%* | 97.83 | 96.03 | 92.88 | 90.06 | 89.10 | 88.56 |
| *20%* | 96.83 | 93.25 | 86.97 | 80.61 | 79.15 | 77.88 |
| *30%* | 95.63 | 90.32 | 80.17 | 71.81 | 68.92 | 67.75 |
| *40%* | 94.21 | 86.55 | 73.71 | 63.94 | 58.68 | 57.10 |
| *50%* | 93.29 | 82.36 | 65.90 | 55.86 | 51.03 | 48.41 |

**Table 1**. Recognition rate ($W_{acc}$) from synthesized speech at the output of the iLBC decoder, 20 ms frame mode.

| Loss Rate | Burst Length | | | | | |
|---|---|---|---|---|---|---|
| | *1* | *2* | *4* | *8* | *12* | *16* |
| *5%* | 98.83 | 98.49 | 97.76 | 96.10 | 95.56 | 95.18 |
| *10%* | 98.69 | 98.16 | 96.43 | 93.23 | 91.89 | 91.16 |
| *20%* | 98.59 | 97.51 | 93.19 | 86.95 | 84.34 | 82.85 |
| *30%* | 98.44 | 96.83 | 89.50 | 80.67 | 76.59 | 74.50 |
| *40%* | 98.30 | 95.71 | 86.17 | 74.57 | 68.19 | 65.81 |
| *50%* | 98.37 | 94.44 | 81.28 | 67.96 | 61.15 | 57.74 |

**Table 2**. Recognition rate ($W_{acc}$) for the iLBC transparametrization technique with the proposed PLC algorithm.

packet, is shown in figure 6. The selected channel conditions are shown in table 3. We can observe that iLBC B-NSR not only outperforms any of the three basic NSR systems, but it also provides a performance close to DSR. This behaviour makes it a serious alternative for speech recognition in IP networks.

| *Channel condition* | *0* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|---|
| *Loss rate (%)* | 0 | 10 | 20 | 30 | 40 | 50 |
| *Burst Length* | – | 1 | 2 | 4 | 8 | 16 |

**Table 3**. Selected channel conditions used in Figure 6.

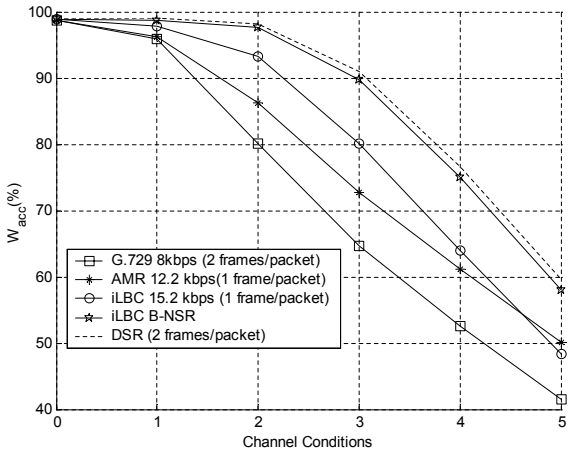


**Fig. 6**. Comparison between recognition rates WAcc for DSR, bitstream-NSR and basic NSR (iLBC, G.729 and AMR).

## 4. SUMMARY

In this work, we have presented a bitstream-based NSR system using the iLBC coder. We have designed an iLBC transparametrization method which transforms the received codec parameters into an MFCC-based feature vector. Also, a packet loss concealment technique, based on interpolation and repetition of parameters, is implemented to mitigate packet losses. The obtained results show a recognition rate close to that of DSR and clearly better than that of the recognition from synthesized speech for all channel conditions.

## 5. REFERENCES


[1] A.M. Peinado, J.C. Segura: "Speech Recognition Over Digital Channels". *Wiley*, 2006.

[2] D. Pearce: "Enabling New Speech Driven Services for Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-ends". *AVIOS 2000: The Speech Applications Conference*, San Jose (USA), 2000.

[3] 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 2002 Workshop on Speech Coding*, December 2002.

[4] D. Pearce and H-G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions", *Proc. ICSLP*, Vol.4, pp.29-32, 2000.

[5] H.K. Kim and R.V. Cox: "A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System". *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, July 2001.

[6] C. Peláez and A. Gallardo and F. Díaz: "Recognizing Voice Over IP: A Robust Front-End for Speech Recognition on the World Wide Web". *IEEE Trans. on Multimedia*, vol. 3, no. 2, June 2001.

[7] A.M. Gómez, A.M. Peinado, V. Sánchez, A.J. Rubio: "Recognition of Coded Speech Transmitted Over Wireless Channels". *IEEE Trans. on Wireless Communications*, vol. 5, no. 9, September 2006.