Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Signal Processing 86 (2006) 1502–1508

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

A comparison of front-ends for bitstream-based ASR over IP

Carmen Peláez-Moreno*, Ascensión Gallardo-Antolín, Diego F. Gómez-Cajas, Fernando Díaz-de-María

Dpto. de Teoría de la Señal y Comunicaciones, EPS-Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911-Leganés, Madrid, Spain

Received 5 February 2004; received in revised form 17 May 2005
Available online 30 September 2005

Abstract

Automatic speech recognition (ASR) is called to play a relevant role in the provision of spoken interfaces for IP-based applications. However, as a consequence of the transit of the speech signal over these particular networks, ASR systems need to face two new challenges: the impoverishment of the speech quality due to the compression needed to fit the channel capacity and the inevitable occurrence of packet losses.

In this framework, bitstream-based approaches that obtain the ASR feature vectors directly from the coded bitstream, avoiding the speech decoding process, have been proposed ([S.H. Choi, H.K. Kim, H.S. Lee, Speech recognition using quantized LSP parameters and their transformations in digital communications, Speech Commun. 30 (4) (2000) 223–233. A. Gallardo-Antolín, C. Pelàez-Moreno, F. Díaz-de-María, Recognizing GSM digital speech, IEEE Trans. Speech Audio Process., to appear. H.K. Kim, R.V. Cox, R.C. Rose, Performance improvement of a bitstream-based front-end for wireless speech recognition in adverse environments, IEEE Trans. Speech Audio Process. 10 (8) (2002) 591–604. C. Peláez-Moreno, A. Gallardo-Antolín, F. Díaz-de-María, Recognizing voice over IP networks: a robust front-end for speech recognition on the WWW, IEEE Trans. Multimedia 3(2) (2001) 209–218], among others) to improve the robustness of ASR systems. LSP (Line Spectral Pairs) are the preferred set of parameters for the description of the speech spectral envelope in most of the modern speech coders. Nevertheless, LSP have proved to be unsuitable for ASR, and they must be transformed into cepstrum-type parameters. In this paper we comparatively evaluate the robustness of the most significant LSP to cepstrum transformations in a simulated VoIP (voice over IP) environment which includes two of the most popular codecs used in that network (G.723.1 and G.729) and several network conditions. In particular, we compare 'pseudocepstrum' [H.K. Kim, S.H. Choi, H.S. Lee, On approximating Line Spectral Frequencies to LPC cepstral coefficients, IEEE Trans. Speech Audio Process. 8 (2) (2000) 195–199], an approximated but straightforward transformation of LSP into LP cepstral coefficients, with a more computationally demanding but exact one. Our results show that pseudocepstrum is preferable when network conditions are good or computational resources low, while the exact procedure is recommended when network conditions become more adverse.

Keywords: Robust speech recognition; Speech coding; IP networks; Coding distortion; Packet loss; LSP

*Corresponding author. Tel.: +34 91 624 8771; fax: +34 91 624 8749.
E-mail addresses: carmen@tsc.uc3m.es (C. Peláez-Moreno), gallardo@tsc.uc3m.es (A. Gallardo-Antolín), dgomez@tsc.uc3m.es (D.F. Gómez-Cajas), fdiaz@tsc.uc3m.es (F. Díaz-de-María).

0165-1684/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2005.08.011

1. Introduction

As voice transmission over IP networks (VoIP) becomes popular, new voice-enabled services provided through these networks are being developed. Therefore, ASR (automatic speech recognition) is called to play an important role in the provision of user-friendly spoken interfaces for these services. However, under those circumstances, two VoIP-specific problems emerge: first, the scarcity of bandwidth makes the use low-to-medium-rate speech coders necessary and, consequently, coding distortion reduces the recognizers accuracy [6,7]; and second, packet losses, severely affect ASR performance [4].

Recent papers ([1–3,6,8], among others) have established that more robust parameterizations can be obtained by transforming some of the parameters sent by the coder, instead of decoding the speech signal and using a conventional ASR front-end. This means that selecting just the necessary information from the bitstream is better than extracting it from the decoded waveform.

The motivations are the following: first, the avoidance (except for quantization) of the encoding-decoding distortion; second, the possibility of selecting just the relevant information for recognition from the bitstream, therefore minimizing the likelihood that the feature extraction process be influenced by irrelevant (from the ASR point of view) or erroneous information (due to channel distortions); and third, the error recovery mechanisms provided by the standard coders can also be improved, adapting them to the ASR problem. This can be achieved by relaxing the restrictions posed by the coding procedures such as maximum delays or light-weight interpolation methods.

Most of the modern speech coders (G.723.1, G.729, and the new AMR—Adaptive Multi Rate—set of coders, for example) employ LSP (Line Spectral Pairs, also called LSF—Line Spectral Frequencies—) parameters for the coding of the speech spectral envelope [9]. There are a number of reasons that motivate this choice: first, they are highly predictable (they give smooth frame to frame transitions); second, their interpretation as frequencies eases the integration of auditory-related concepts; finally, they offer the possibility of performing a straight-forward stability check. However, the use of LSP as feature vectors has proved to be unsuitable for current ASR systems [1]. Therefore, they must be transformed into MFCC-type (mel frequency cepstral coefficients) parameters, which nowadays are still the most successful parameters for ASR.

Since bitstream-based ASR front-ends turn out to be more robust for dealing with compressed speech, and current coders use LSPs parameters for representing the speech spectral envelope, the study of the robustness of the transformation methods for obtaining MFCC-type parameters from LSPs becomes relevant. Thus, in this paper we conduct a comparative evaluation of alternative computation methods to obtain mel-scaled LPCC (linear prediction cepstral coefficients), i.e., the calculation of MFCC from LP-based parameters (LSPs in our case).

On the one hand, a proposal by Kim et al. [1,5] called pseudocepstrum provides a straight and computationally efficient approximation to the LPCC parameterization. On the other, the LPCC parameterization can be computed in an exact and computationally more demanding way. Both approaches have been compared by Kim et al. proving comparable performances considering the quantization errors introduced by a speech codec. In this paper, we compare their robustness to both the speech coding stage distortion and the impairments due to the IP transmission channel. In particular, we have tested both parameterizations in several simulated VoIP scenarios using two codecs (G.723.1 and G.729) and a wide range of packet loss rates (PLRs) and mean burst lengths (MBLs). This realistic testing environment adds to the analysis of the LSP quantization effects of [1] an evaluation of the influences of the whole encoding-decoding process (for example on the energy parameter extraction or the frame rate provided) and the network distortions. This allows us to discuss when the approximation given by pseudocepstrum is advantageous and when, on the contrary, the exact LSP to MFCC conversion is preferable.

Finally, though not considered in this work, it is worth mentioning an alternative method for avoiding both the coding and decoding stage called Distributed Speech Recognition (DSR), which consists of a standard protocol for sending a specific type of ASR parameterization extracted at the user-end, instead of the coded version of the whole speech signal [10]. This is a very convenient alternative in terms of ASR performance, requiring only that the user terminal implements the standard parameterization defined in the DSR protocol.