Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 58 (2014) 111–123

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Low bit rate compression methods of feature vectors for distributed speech recognition

Jose Enrique Garcia *, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

Communications Technology Group (GTC), Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain

Received 28 September 2012; received in revised form 3 October 2013; accepted 16 November 2013
Available online 28 November 2013

Abstract

In this paper, we present a family of compression methods based on differential vector quantization (DVQ) for encoding Mel frequency cepstral coefficients (MFCC) in distributed speech recognition (DSR) applications. The proposed techniques benefit from the existence of temporal correlation across consecutive MFCC frames as well as the presence of intra-frame redundancy. We present DVQ schemes based on linear prediction and non-linear methods with multi-layer perceptrons (MLP). In addition to this, we propose the use of a multipath search coding strategy based on the *M*-algorithm that obtains the sequence of centroids that minimize the quantization error globally instead of selecting the centroids that minimize the quantization error locally in a frame by frame basis. We have evaluated the performance of the proposed methods for two different tasks. On the one hand, two small-size vocabulary databases, Spechdat-Car and Aurora 2, have been considered obtaining negligible degradation in terms of Word Accuracy (around 1%) compared to the unquantized scheme for bit-rates as low as 0.5 kbps. On the other hand, for a large vocabulary task (Aurora 4), the proposed method achieves a WER comparable to the unquantized scheme only with 1.6 kbps. Moreover, we propose a combined scheme (differential/non-differential) that allows the system to present the same sensitivity to transmission errors than previous multi-frame coding proposals for DSR.

Keywords: Distributed speech recognition; Neural Networks; Multi-layer perceptron; Predictive vector quantizer optimization; IP networks

1. Introduction

Implementing automatic speech recognition (ASR) applications in mobile devices can be extremely inconvenient due to its computational and memory constraints. Distributed speech recognition (DSR) allows ASR applications to be used in mobile devices by distributing the ASR system between the front-end at the client side and the back–end at the server side (Pearce, 2000, 2004). The front-end module extracts and compresses the acoustic features and then transmits them to the back–end. There the decompression and recognition algorithms run.

Nevertheless, not only mobile devices can benefit from the use of DSR. DSR is also convenient for other applications like speech enabled web browsing, to reduce the computational load at the client side (Digalakis et al., 1998). As an alternative to DSR, network speech recognition (NSR) sends coded speech. From a practical point of view, NSR can benefit from the existence of a large number of applications using VoIP. Moreover, the decoded speech signal can be stored and used again in the future for improving the ASR system (retrain acoustic models, use different acoustic features,...). However, several studies show that the performance of NSR in terms of word error rate (WER) is dramatically reduced using state of the art speech codecs at low bit-rate conditions (Digalakis et al., 1998; Kelleher et al., 2002; Kiss, 2000; Srinivasamurthy et al., 2006). The main reason for this is that most speech coding algorithms are designed for maximizing speech perceptual quality

* Corresponding author. Address: Electronic Engineering and Communications Department, University of Zaragoza, C/ Maria de Luna 1, 50018 Zaragoza, Spain. Tel.: +34 625837839.
E-mail addresses: jegarlai@unizar.es (J.E. Garcia), ortega@unizar.es (A. Ortega), amiguel@unizar.es (A. Miguel), lleida@unizar.es (E. Lleida).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.11.007

which does not necessarily result in optimal recognition performance.

In order to reduce the resources needed by the system, efficient low bit-rate compression methods are desired for DSR. Although nowadays wireless networks allow high bit-rate transmissions, in some applications, the server must support a high number of clients sharing the available bandwidth and, sometimes, mobile device users have contracts in which the Internet Service Provider charges them on a usage basis (per amount of data transferred). Under those circumstances, a reduced bandwidth would result in clear benefits for the final user.

One of the most extended acoustic representations for ASR are Mel frequency cepstral coefficients (MFCC) and so far, several feature compression schemes have been proposed in the literature. In Digalakis et al. (1998, 1999), the authors evaluate several approaches for compressing MFCCs. After analyzing uniform and non-uniform scalar quantization along with vector quantization for bit-rates ranging from 1.2 kbps to 10.4 kbps, the most efficient approach was split vector quantization. The proposal in Ramaswamy and Gopalakrishnan (1998) makes use of a one-step, scalar, linear predictor at a fixed rate of 4.0 kbps achieving similar recognition performance than the uncompressed features for several languages.

Representation of speech by means of MFCC presents inter and intra-frame redundancy that can be exploited to reduce the bit-rate as much as possible. In Srinivasamurthy et al. (2006) inter-frame correlation is exploited by means of a DPCM approach, but uniform scalar quantization is used disregarding MFCC intra-frame correlation. In order to benefit from both sources of redundancy, intra and inter-frame, So and Paliwal (2006) proposed a scalable fixed-rate multiframe GMM-based block quantization scheme that outperforms the previous approaches in terms of bandwidth savings for a small vocabulary task such as Aurora-2.

In this work, we present an ensemble of algorithms that benefit from inter and intra-frame redundancy to compress MFCC features for DSR. The proposed encoders belong to the family of predictive or differential vector quantizers, that uses signal prediction techniques and try to exploit temporal correlation between adjacent frames. This temporal correlation is due to the overlapping of the windowing step and the relatively slow variation of speech production. In addition to this, these methods exploit mutual information between feature vector coefficients in the same frame thanks to vector quantization (VQ). Two main problems arise when using only one codebook for all the coefficients in a frame: long time for training the VQ and high computation and storage requirements at the client side for the coding stage. Although the former is not a major problem since VQ must be trained only once and this training is always performed offline, the latter can make the task not affordable in real time for low-resource devices. To reduce the computational resources needed at the client side, we adopt the solution proposed in Digalakis et al. (1999), split VQ. Our first proposal (Garcia et al., 2009) makes use of a first-order linear predictor that performs vector quantization of prediction errors. We present here a deeper study on the DVQ coding scheme along with new advances that enhance its performance. The second compression scheme makes use of a non-linear function by means of a multilayer perceptron (MLP) that has an input layer with the latest quantized coefficients along with energy information as the one presented in Garcia et al. (2010). Prediction gain increases and thus the same WER can be obtained with lower bit-rates. In this work, we additionally propose a global optimization approach to improve the suboptimal performance of single-path search methods for predictive vector quantization. Instead of choosing the closest codeword in a frame by frame basis, disregarding the influence of these decisions on future predictions, a global optimization strategy can be followed. By means of a delayed decision coding approach, using the M-algorithm (Jelinek and Anderson, 1971), the best M quantization hypotheses are kept every time a new frame arrives. Once the final time index is reached, the hypothesis with the minimum accumulated squared error provides the sequence of values to be sent to the decoder. In the experiments carried out in this work, all the proposed methods outperform the ETSI standard compressor and typical VQ approaches in a connected digits task and in the 5 kword large vocabulary task of Aurora 4 (Hirsch, 2001), for transmission rates considerably lower.

The issue of channel error robustness for DSR is a large research topic that has attracted interest from many researchers in the last years. Packet loss in DSR is a challenging problem that can degrade the performance of the system. However, the aim of this paper is not to analyze the influence of transmission errors in DSR but to propose a novel compression scheme. When several frames are grouped together to be sent to the server side, a combined approach can be followed with static and predictive coefficients achieving the same robustness against packet losses than other approaches previously proposed (Srinivasamurthy et al., 2006; So and Paliwal, 2006). Moreover, the same robustness techniques that have been proposed in the literature can be used with the methods we present here in order to reduce the transmission error degradation. In Tan et al. (2005) the authors classify the robustness techniques for DSR into three different categories: error detection, error recovery and error concealment. Client-based techniques like retransmission, interleaving and forward error correction (FEC) are able to recover a large amount of transmission errors (Bernad and Alwan, 2002; James and Milner, 2004; Peinado et al., 2005; Milner and James, 2006; Gomez et al., 2006; Tan et al., 2007; Gomez et al., 2009; Flynn and Jones, 2010, 2012). In combination with these techniques, server-based error concealment exploits the redundancy of speech by means of feature-reconstruction (Boulis et al., 2002; Milner and Semmani, 2000; Tan et al., 2004; Gomez et al., 2004; James