Computers and Electrical Engineering 51 (2016) 12–25

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

An efficient low bit-rate compression scheme of acoustic features for distributed speech recognition☆

Azzedine Touazi[a,b,*], Mohamed Debyeche[a]

[a] Speech Communication and Signal Processing Laboratory (LCPTS), Faculty of Electronics and Computer Science, University of Science and Technology Houari Boumediene (USTHB), Bab Ezzouar, Algiers, Algeria
[b] Center for Development of Advanced Technologies (CDTA), Baba Hassen, Algiers, Algeria

ARTICLE INFO

Article history:
Received 24 September 2015
Revised 22 February 2016
Accepted 22 February 2016
Available online 10 March 2016

Keywords:
Distributed speech recognition
Weighted least squares fitting
Aurora-2 database
Low bit-rate source coding

ABSTRACT

Due to the limited network bandwidth, a noise robust low bit-rate compression scheme of Mel frequency cepstral coefficients (MFCCs) is desired for distributed speech recognition (DSR) services. In this paper, we present an efficient MFCCs compression method based on weighted least squares (W-LS) polynomial approximation through the exploitation of the high correlation across consecutive MFCC frames. Polynomial coefficients are quantized by designing a tree structured vector quantization (TSVQ) based scheme. Recognition experiments are conducted on the noisy Aurora-2 database, under both clean and multi-condition training modes. The results show that the proposed W-LS encoder slightly exceeds the ETSI advanced front-end (ETSI-AFE) baseline system for bit-rates ranging from 1400 bps to 1925 bps under clean training mode. However, a negligible degradation is observed in case of multi-condition training mode (around 0.6% and 0.2% at 1400 bps and 1925 bps, respectively). Furthermore, the obtained performance generally outperforms the ETSI-AFE source encoder at 4400 bps under clean training and provides similar performance, at 1925 bps, under multi-condition training.

1. Introduction

During the last few years, the implementation of client-server architecture has received more attention for the practical speech recognition systems, especially for mobile applications. In the client-server speech recognition, also known as distributed speech recognition (DSR) [1], the front-end client is included in the terminal and it is connected over a data channel to a remote recognition server (back-end). DSR provides particular benefits for mobile terminal services such as giving access from different points of network with a guaranteed level of recognition performance. The Mel frequency cepstral coefficients (MFCCs) are the commonly used feature components for DSR front-ends [2–4]. These features are extracted and quantized at the client side, and then transmitted through an error protected data channel to a hidden Markov model-based (HMM) speech recognition system.

The introduction of new mobile services sometimes tends to produce new saturations in the network, as the available channel bandwidth is relatively limited. One solution is to quantize the feature vectors using the least amount of bits, which

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Associate Editor Dr. Z. Arnavut.

* Corresponding author at: Speech Communication and Signal Processing Laboratory (LCPTS), Faculty of Electronics and Computer Science, University of Science and Technology Houari Boumediene (USTHB), Bab Ezzouar, Algiers, Algeria. Tel.: +213 5563360093.
E-mail addresses: atouazi@cdta.dz, touazi.azzedine@gmail.com (A. Touazi), mdebyeche@gmail.com (M. Debyeche).

http://dx.doi.org/10.1016/j.compeleceng.2016.02.019
0045-7906/

can be supported by the available channel bandwidth, while keeping the recognition performance that is as close as possible to that of unquantized feature vectors. In fact, several techniques for compressing MFCCs, in DSR systems, have been designed. Most of the state of the art approaches exploit inter and/or intra-frame correlations across consecutive MFCC components. This offers the capability to efficiently design low bit-rate source coding schemes. Among these methods, one can cite the work in Ref. [5], where eight temporally consecutive 14-dimensional MFCC vectors are grouped and then processed by the discrete cosine transform (DCT). The achieved compression bit-rate is around 4200 bps. In the same manner as for one-dimensional DCT, a two-dimensional DCT (2D-DCT) has been addressed in Ref. [6] where both inter and intra-frame correlations are exploited. In this method, for each 12×12 block of consecutive MFCC frames a 2D-DCT is applied and only the DCT components with the highest energy are quantized while the rest of components are set to zero. No significant performance degradation is obtained with bit-rates as low as 624 bps for speaker dependent isolated digit recognition.

The European telecommunication standards institute (ETSI) [2–4] defines split vector quantization (SVQ) scheme [7], where MFCC coefficients are grouped into pairs and each pair is quantized using its own vector quantization (VQ) codebook. The resulting MFCC encoding bit-rate is 4400 bps without including channel error protection.

A novel bits allocation scheme for ETSI front-end (FE) [2] has been successfully applied in Ref. [8]. The quantization bits are allocated proportionally to the mutual information measures between FE sub-vectors, where the greater portion of the total bit-rate is assigned to the lower MFCCs. This method has yielded significant performance improvements for clean speech data. The authors in Ref. [9] presented a scalable predictive approach in which each feature is independently quantized using a scalar predictive coding. The scalability allows providing flexibility in optimizing the DSR bit-rate, in terms of recognition performance, to the changing bandwidth requirement and server load.

The half frame rate (HFR) front-end algorithm [10,11] investigates the redundancies in the full frame rate (FFR) features of ETSI-FE, where the source coding bit-rate is reduced from ETSI-FE 4400 bps to 2200 bps. The HFR algorithm has been evaluated on the Aurora-2 [12] clean speech. The comparison of achieved performance accuracy levels are close to ETSI-FE compression algorithm. Another DSR encoder has been proposed in Ref. [13], called packetization and variable bit-rate compression scheme. This encoder has the property of being compatible with various VQ-based DSR encoders. The coded MFCC frames are grouped using the group of pictures (GoPs) structure taken from video coding, and then a Huffman coding is applied for each group. The packetization and variable bit-rate method provides lossless compression at 3400 bps for Aurora-2 clean data, whereas the GoP grouping of ETSI-FE coded frames requires an additional algorithmic delay.

Moreover, a series of quantization techniques have been described in Ref. [14], where both inter and intra-frame MFCC correlations are exploited. One of the most discussed methods is the multi-frame Gaussian mixture model-based (GMM) block quantizer [15]. Evaluated on Aurora-2 clean speech, the GMM-based encoder achieved the best recognition performance at lower bit-rates, exhibiting a negligible 1% degradation at 800 bps. The GMM-based method has been extended to quantize MFCCs in noisy environments [16]; however, the obtained results showed a degraded recognition performance by increasing the noise level.

More recently, the authors in Ref. [17] have proposed a new bandwidth reduction scheme based on Haar wavelet decomposition. Experiments are performed on Aurora-2 noisy speech under clean training condition. Compared with the baseline system, when there is no packet loss (i.e. source coding), the bandwidth can be reduced to 50% without degrading the recognition performance. However, a graceful degradation is obtained when the bandwidth is reduced to 25% of the baseline. In addition, the work in Ref. [18] presents a series of low bit-rate quantization methods based on differential vector quantization (DVQ) algorithms. The performance is evaluated for two different tasks (Aurora-2 and Aurora 4 [19] for small and large vocabulary tasks, respectively) using only clean speech. Results obtained show that the DVQ-based schemes can be an efficient compression method at very low bit-rate, in particular for small vocabulary DSR applications. Generally speaking, most of the previously proposed low bit-rate DSR encoders suffer from degraded performance under noisy conditions.

The method we propose in this paper focuses on reducing the source coding bit-rate of MFCC vectors in a DSR system, using weighted least squares approximation. According to the DSR limitations in terms of bandwidth, memory, and computational requirements, our ultimate aim is twofold: (i) the compression task should not cause any significant loss on recognition performance, in particular under noisy conditions and (ii) the computational complexity and memory requirements should be moderate. The key idea behind this method is that we do not have to transmit every set of extracted MFCCs to the decoder (back-end); instead, we could transmit only the coefficients of the polynomial that approximates these MFCCs. At the server side, the MFCC components are reconstructed using the de-quantized polynomial coefficients. However, the performance will depend not only on the amount of allocated bits but also on both polynomial degree and weighting values.

A set of temporally consecutive MFCC frames are extracted from the speech utterance and grouped into blocks, where each block row corresponds to the time trajectory of a particular cepstral feature. By means of exploiting both the slow evolution and correlation characteristics across MFCC frames for dimensionality reduction purpose, each block row is approximated by low degree polynomial, through a weighted least-squares sense. The method used to calculate the weighting coefficient of each block column is inspired from the variable frame rate (VFR) algorithm proposed in Ref. [20]. The calculation of weights is based on the log energy parameter in which the larger weight is assigned to the more noise robust MFCC frame. Furthermore, *QR* factorization is used for solving the weighted least-squares problem.

In earlier work, we introduced the idea of applying the unweighted least squares (U-LS) approximation (i.e. all features have the same weight) to encode MFCCs [21]. The promising initial results obtained have shown that the approach could be further explored. Here, we present an extension to the U-LS-based encoder at lower bit-rates through the introduction