DISTRIBUTED SPEECH RECOGNITION WITH CODEC PARAMETERS

Bhiksha Raj[1], *Joshua Migdal*[2], *and Rita Singh*[3]
1. Mitsubishi Electric Research Labs, Cambridge, MA 02139 USA
2. Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA
3. Carnegie Mellon University, Pittsburgh, PA 15213

ABSTRACT

Communication devices which perform distributed speech recognition (DSR) tasks currently transmit standardized coded parameters of speech signals. Recognition features are extracted from signals reconstructed using these on a remote server. Since reconstruction losses degrade recognition performance, proposals are being considered to standardize DSR-codecs which derive recognition features, to be transmitted and used directly for recognition. However, such a codec must be embedded on the transmitting device, alongwith its current standard codec. Performing recognition using codec bitstreams avoids these complications: no additional feature-extraction mechanism is required on the device, and there are no reconstruction losses on the server. In this paper we propose an LDA-based method for extracting optimal feature sets from codec bitstreams and demonstrate that features so derived result in improved recognition performance for the LPC, GSM and CELP codecs. For GSM and CELP, we show that the performance is comparable to that with uncoded speech and standard DSR-codec features.

1. INTRODUCTION

Cellphones and PDAs have lately become very popular and are being used for multiple tasks, which sometimes require complex and involved instructions. These devices are typically small in size and it is usually very inconvenient and inefficient to use the inbuilt input devices provided to feed in complex command sequences. In this respect, speech is a very convenient and natural interface. The development of good speech interfaces has lately motivated a lot of research. The main problem faced in these efforts is that while it is feasible to incorporate speech recognition systems within these devices, their size limits the complexity of the task that they can handle. This is because more complex tasks typically involve more complex grammars, larger vocabularies, parsing mechanisms, etc. It is therefore considered to be more practical and efficient to perform the recognition, with subsequent task completion, on a remote server

. Currently this is accomplished by transmitting the coded speech to the server, which reconstructs the speech signal, parametrizes it and performs recognition. However, it is well known that speech which has undergone coding and reconstruction results in lower recognition accuracies than uncoded speech [1]. To avoid this, a reasonable solution that has been proposed is to extract recognition features from the speech signal and transmit those instead of the codec parameters. The server can then use these features directly for recognition. Since the speech has not undergone coding and decoding in this process, there are no additional coding related losses incurred in the recognition. This scheme of distributing the speech recognition task between the transmitting handset and the remote device is referred to as *distributed speech recognition* (DSR).

When this kind of DSR is an add-on on devices such as cellphone handsets or PDAs which function in the conventional manner, the device must actually incorporate a codec that can compute the recognition features. In addition, protocols must be established to distinguish between when the transmitter is transmitting regular codec parameters for decoding to speech and when it is transmitting recognition features. This necessitates the establishment of universal standards for such codecs and protocols in order for any cellphone or PDA to be able to communicate with any speech recognition server. Standards bodies such as the European Telecommunication Standards Institute (ETSI) and the International Telecommunication Union (ITU) are currently in the process of defining such standards.

There are still some hurdles in the composition of such standards: firstly they must be designed to accommodate a fast changing technology, and secondly, the various cellphone and PDA manufacturers and the telephony providers must be convinced to make appropriate product adjustments to conform to these standards.

The requirements would however be simplified if the devices could continue to simply transmit coded speech parameters, but if recognition features could be derived directly from these. This would eliminate losses incurred due to further reconstruction of speech from the coded parameters. This would also eliminate the need for the transmitting device to incorporate an alternate codec.

This alternative approach to DSR, where the recognition features are computed directly from the codec parameters transmitted by standard codec, has been proposed earlier by several researchers [2-5]. In all of these bitstream-based feature extraction methods, however, the optimal combination of features derived from the short-term (LPC) and long-term (residual) components of the bitstreams was obtained either through exhaustive experimentation [4] or heuristically [5]. In general, the performance achieved, while superior to that obtained with decoded speech, has been inferior to that obtained with uncoded speech.

In this paper, in addition to presenting further evidence to show that bitstream-based feature extraction is a viable alternative to having alternate codecs in the transmitting device, we propose an LDA-based scheme for optimal combination of information derived from the LPC and residual components of the bitstream to construct features for recognition. We present results for three different coding schemes, GSM, CELP and LPC, where we show that the features so derived can not only result in better recognition accuracies than those obtained with the decoded (or reconstructed) speech, but also, in the case of medium and high-bitrate codecs like GSM and CELP, result in recognition accuracies comparable with those obtained with uncoded speech.

In Section 2 of this paper we describe the WI007 front end specified by ETSI for DSR. This front end was designed for use in cases where recognition features are to be computed on the device and transmitted subsequently [7]. WI007 has therefore been used in this paper to perform recognition experiments with the uncoded speech, in order to establish a baseline. In experiments which involve decoded or reconstructed speech, MFCCs have been used as features. In Section 3 we describe the GSM, CELP and LPC codecs evaluated in this paper. In Section 4 we describe how recognition features are derived from the long-term and short-term components respectively of the bitstreams of each of these codecs. In Section 5 we describe the LDA-based procedure used to derive the final bit-stream based recognition features in each case. In Section 6 we provide our experimental results. Finally, in Section 7 we present our conclusions.

2. THE WI-007 STANDARD FRONT END

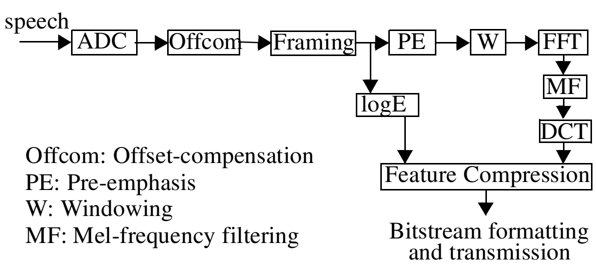

Figure 1. Block diagram of the ETSI-Aurora WI-007 frontend codec (source [7]).

The WI-007 codec is a distributed speech recognition front end specified by the European Telecommunications Standards Institute (ETSI), for use on cellular phones and other communication devices that connect to speech recognition servers [7]. A block diagram of the WI007 front end is given in Fig. 1. Input speech (sampled at 8khz for experiments reported in this paper) is first subjected to DC offset removal using a notch filter. The signal is windowed into frames of 25ms in length, with adjacent frames overlapping by 15ms. The frames are preemphasized and smoothed using a Hamming window, then subjected to a fast Fourier transform. 23 Mel-frequency spectral terms covering the frequency range 64Hz-4000Hz are derived from them. The logarithm of the Mel-frequency spectra are passed through a discrete cosine transform to derive 13-dimensional Mel-frequency cepstral coefficients.

The cepstral vectors thus obtained are further compressed for transmission. Beginning with the second cepstral component, pairs of cepstral components are vector quantized using codebooks with 64 components. The first component of the cepstral vectors is paired with the log energy of the frame, and the pair is quantized using a 256 component codebook. The transmitted features have an bitrate of 4800 bits per second.

In our experiments, we deviate from the exact specifications for this codec in that we do not quantize the cepstral component pairs. We expect to get slightly better baselines because there are no quantization losses involved. In establishing a baseline we also assume that no bit errors are introduced during transmission.

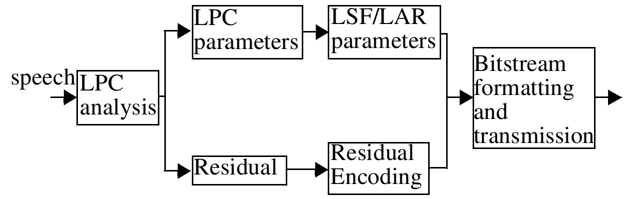

Figure 2. Block diagram of typical linear predictive codec. LPC analysis is followed by parameterization and encoding of LPC and residual parameters, which are then formatted for transmission.

3. CODING SCHEMES

All codecs considered in this paper are linear predictive coding (LPC)-based codecs. In LPC-based codecs, frames of speech, typically between 20 and 30ms long, are decomposed into a linear prediction filter and an excitation signal, called the residual, using LPC analysis. The LPC parameters and the residual are further coded and transmitted. The primary difference between these coding schemes is in the manner in which the residual is coded, although they also vary in the size of the analysis window, the order of LPC performed, and the manner in which LPC parameters are coded. In this paper we have specifically considered three such codecs: GSM, CELP and LPC.

3.1. The GSM fullrate codec

The GSM codec is a linear predictive coder that uses Regular Pulse Excitation, Long Term Prediction (RPE-LTP) to encode a speech signal. The GSM codec encodes 160-sample (20ms) frames of preprocessed, 13-bit PCM speech, sampled at a rate of 8kHz, into RPE-LTP quantized parameters using 260 bits, resulting in an overall bitrate of 13 kilobits per second.

Preprocessing is done on a per-frame basis. Each frame is first subjected to a DC offset compensation filter and then to a first order FIR preemphasis filter with a preemphasis factor of $28180/2^{15}$. LPC analysis is performed on each frame, and 8^{th} order LPC reflection coefficients are derived. The reflection coefficients are transformed to log area ratios and quantized for transmission. A long-term prediction filter, characterized by a long-term gain and a delay, is derived 4 times in each frame, using sub-frames of 40 samples (5ms) each, from the residual of the LPC filter. The residual of the long-term prediction filter within each subframe is then represented by 1 of 4 candidate sequences of 13 samples each. The quantized log area ratios, the long-term delay and gain, and the coded long-term residuals are all transmitted in the GSM bitstream.

3.2. The CELP FS1016 codec

The CELP FS1016 codec is a linear predictive coder that uses Codebook Excited Linear Prediction to encode a speech signal. The CELP codec encodes 240-sample (30ms) frames of 8KHz sampled speech into 144 bits of CELP coded parameters, resulting in an overall bitrate of 4800 bits per second.

Each 240-sample frame of incoming speech is band-pass filtered between 100Hz and 3600Hz and 10^{th} order LPC analysis is performed. The derived LPC coefficients are converted to line spectral frequency (LSF) parameters which are quantized for transmission. The analysis window is further divided into four sub-frames of 60 samples (7.5ms). Within each sub-frame the LPC residual is represented as the sum of scaled codeword entries, one from a fixed codebook and a second from an adaptive codebook that is constructed from the current residual using information about the pitch. The fixed codebook entry is determined using an analysis-by-synthesis approach that minimizes the perceptually weighted error between the original speech signal and the re-synthesized signal. The LSF parameters, the codebook indices and gains, and pitch and gain information required by the adaptive codeword are transmitted.

3.3. The DOD LPC FS1015 CODEC

The LPC FS1015 codec uses linear predictive coding to encode the speech signal. The codec encodes 180-sample (22.5ms) frames of 8KHz sampled speech into 54 bits of LPC coded parameters, resulting in an overall bitrate of only 2400 bits per second.

Each 180 sample (22.5 ms) frame of incoming speech is preemphasized and a 10^{th} order LPC analysis is performed. LPC parameters are transformed to log area ratios for transmission. The residual is modelled either by white noise or by a periodic sequence of pulses, depending on whether the speech frame is identified as being unvoiced or voiced. The log area ratios, the voiced/unvoiced flag, the pitch, and the gain of the LPC filter are transmitted.

4. DERIVING FEATURES FROM BITSTREAMS

Since the codecs described are all LP codecs, the bitstream carries both coded LPC parameters and information from the residual (see Fig. 2.). Recognition parameters can be derived directly from both these components of the bitstreams. The LPC parameters represent the gross spectral characteristics of the speech signal, which are usually the most important characteristics needed for recognizing the speech. The residual, on the other hand, typically captures information relating to the speaker, such as the pitch, and the perceptual quality of the reconstructed signal. Nevertheless, the residual continues to contain information relating to the identity of the underlying speech sounds, and it is important to capture these effectively as well.

Previous attempts at deriving recognition features from bitstreams have either concentrated on deriving spectral information from the LPC component of the bitstream, extracting only energy related information from the residual [2] [3], or have depended on empirically determined combination of features derived from the LPC parameters and the residuals [4] [5].

In our work we combine parameters derived from both the LPC and the residual components of the bitstream in a principled manner to optimize classification performance. First LPC parameters, either LAR parameters (for GSM or LPC) or LSF parameters (for CELP) are extracted from the bitstream. The extracted parameters are interpolated to effectively obtain one set of LPC parameters every 10ms. Cepstral vectors are then derived from these LPC parameters. The excitation signal is also extracted from the bitstream by setting the short-term prediction coefficients to zero and decoding the bitstream. Since it is unclear as to exactly which components of the excitation contain information about the underlying sounds, the entire spectrum of

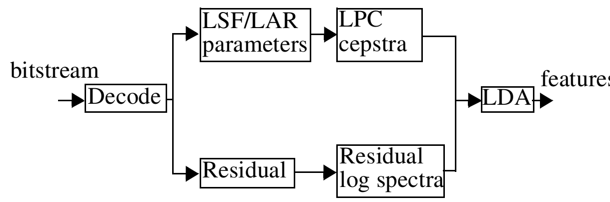

Figure 3. Block diagram showing the LDA-based procedure for extracting recognition parameters from the bitstream.

the excitation signal is analyzed for useful components. 32-dimensional log spectra are derived from the excitation for each frame corresponding to an LPC vector in the interpolated sequence of LPC vectors. An extended vector is formed by concatenating every LPC cepstral vector with the corresponding log-spectral components from the excitation. The dimensionality of the vectors so constructed is then reduced to 13 by performing linear discriminant analysis (LDA) [6] on the extended vectors. The classes that are used for the LDA analysis are the same as the phones modelled by the recognizer. Figure 3. illustrates the complete procedure.

GSM: For the GSM codec 13 dimensional cepstral coefficients are derived from the 8^{th} order LPC coefficients. Every cepstral vector represented 20ms of speech, maintaining synchronicity with the coded bitstream.

CELP: For the CELP codec 15 dimensional cepstral coefficients are derived from the 10^{th} order LPC coefficients in the bitstream. Although the transmitted coefficients represent 30 ms of speech, they are resampled using linear interpolation to represent only 25 ms of speech.

LPC: For the LPC codec 15 dimensional cepstral coefficients are derived from the 10^{th} order LPC coefficients in the bitstream. Each cepstral vector now represents 22.5ms of speech.

5. RECOGNITION EXPERIMENTS

Experiments were conducted using the Resource Management database. The training data consisted of 2800 utterances from the RM database, the test set consisted of 1600 utterances. The vocabulary was 997 words. The dictionary used was the CMU-dict. The CMU SPHINX-III system was used in all experiments. In all experiments triphonetic acoustic models with 2000 tied states were trained, where each state distribution was modelled by a mixture of 4 Gaussians. Only cepstra and difference cepstra were used. A simple bigram LM was used in all experiments. The language weight used was very low in order to emphasize the effect of the acoustic component of the likelihoods on recognition.

Experiments were performed using the GSM, CELP and LPC codecs. For our experiments the GSM 06.10 codec downloadable from Technischen Unviersitaet Berlin [8] was used.. This is a 13kbps codec and is also currently part of the linux suite. The CELP3.3 version LPC10-e version 55 downloadable from the DDVPC website [9] were used for the experiments.

A common baseline was first established by performing recognition on uncoded speech using the WI-007 frontend described in Section 2. This is the recognition accuracy that would be

Coding Scheme	Recognition Feature Set	WER(%)
None	WI-007	12.0
GSM	MFCC from decoded speech	13.3
GSM	LPCcep + c[0] from residual	15.7
GSM	LPCcep + residual logspec + LDA	11.8
CELP	MFCC from decoded speech	15.3
CELP	LPCcep + c[0] from residual	14.6
CELP	LPCcep + residual logspec + LDA	11.4
LPC	MFCC from decoded speech	23.8
LPC	LPCcep + c[0] from residual	24.4
LPC	LPCcep + residual logspec + LDA	21.1

Table 1: Recognition performance obtained with various features derived from speech coded using various codes.

expected, were a communications device to incorporate the WI-007 frontend codec in the absence of quantization and transmission errors. On clean speech this might be considered an upper bound on the performance to be obtained with 8Khz sampled speech using DSR. An additional baseline was established by performing recognition experiments using the coded/reconstructed speech for each of the codecs. Recognition was also performed in each case on the LPC cepstra, augmented with energy information from the residual in the form of the c[0] component of the cepstra. This is the recognition performance obtained using a feature extraction mechanism that does not use information from the residual. The final recognition experiment in each case was performed using 13 dimensional feature vectors derived from the bitstream using the LDA procedure described earlier. Recognition results from these experiments are reported in Table 1.

From Table 1, we observe that the word error rates obtained using the reduced-dimensionality bitstream feature are better than the performance obtained by recognizing encoded/decoded speech. In fact, for the fullrate GSM codec and CELP they are slightly better than the recognition performance obtained with WI-007. While the recognition performance with the LPC codec is very poor, the LDA-based feature results in an improvement in performance nevertheless. It must also be noted that any comparison of the LPC codec with WI-007 would not be completely fair as the latter has twice the bitrate of the former.

6. CONCLUSIONS

In this paper we have proposed an LDA-based method for deriving optimal feature sets from bitstreams of encoded speech in the case of GSM, LPC and CELP codecs. The experiments we have reported show that it is indeed possible to obtain recognition performance that is comparable with, if not better than, that obtained with uncoded speech using features derived directly from the bitstreams of these codecs. Thus it is feasible to design DSR systems where feature derivation need not be performed on a user's handheld device, reducing the importance of changing existing coding and transmission standards. Nevertheless, there remain several advantages to having the communications device transmit front-end features. Such a front-end codec could, in principle, parameterize full-bandwidth speech that has been sampled at bitrates greater than 8000Hz, which would result in much greater recognition accuracies. Bitstream-based feature representations provide an intermediate route where much better recognition accuracies can be obtained than with the decoded speech using traditional communications devices which do not incorporate the front-end codecs or the transmission protocols that go with them. The LDA-based method proposed in this paper furthers this end by presenting an automated mechanism for deriving optimal representations from bitstreams.

7. REFERENCES

1. Lilly, B.T., and Paliwal, K.K., (1996) “Effect of speech coders on speech recognition performance”, Proc. ICSLP 1996

2. Choi, S.H., Kim H.K, Lee, H.S, and Gray, R.M, (1998) “Speech recognition method using quantised LSP parameters in CELP-type coders”, Electron. Lett., vol 34, no. 2, pp. 156-157, Jan 1998

3. Gallardo-Antolin, A., Diaz-de-Maria, F., and Valverde-Albacete, F. (1998), “Recognition from GSM digital signal”, Proc. ICSLP 1998.

4. Huerta, H. and Stern, R.M., (1998) “Speech Recognition from GSM codec parameters”, Proc. ICSLP 1998.

5. Kim, H.K., and Cox, R. (2000), “Bitstream-based feature extraction for wireless speech recognition”, Proc. ICASSP 2000

6. Duda, R.O, Hart, P. E., and Stork, D.G., (2001) *Pattern Classification*, John Wiley and Sons Inc., New York, NY

7. ETSI (2000), “Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithms” European Telecommunications Standards Institute, Document ETSI ES201 108 V1.1.2 (2000-04), http://www.etsi.org

8. http://kbs.cs.tu-berlin.de

9. http://www.plh.af.mil/ddvpc/index.html