SIGNAL PROCESSING BASED SEGMENTATION AND HMM BASED ACOUSTIC CLUSTERING OF SYLLABLE SEGMENTS FOR LOW BIT RATE SEGMENT VOCODER AT 1.4 KBPS

Sadhana Chevireddy, Hema A. Murthy, C. Chandra Sekhar

Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai, 600036, INDIA
email: sadhana@cse.iitm.ernet.in, hema@cse.iitm.ac.in, chandra@cse.iitm.ac.in

ABSTRACT

In this paper, we propose a novel approach for developing a segment-based vocoder at very low bit-rates. The segmental unit chosen for coding is a syllable. A signal processing technique called automatic group delay based segmentation is used to obtain syllable like segments. The segment codebook is prepared by acoustically clustering the syllable segments using a Hidden Markov Model (HMM) based unsupervised and incremental training algorithm. When the residual is modeled using MELP, a bit-rate of 1.4 Kbps is achieved. The synthesized speech quality is compared with that of the standard MELP codec at 2.4 Kbps using the objective evaluation measure, PESQ.

1. INTRODUCTION

The goal of speech coding is to preserve the intelligibility at very low bit rates. To enable this, very low bit-rate coders are developed. Most very low bit-rate coders use a segment-based approach to coding [1] [2] [3].

The *first* phase in building such a coder is the segmentation task. Here, large speech corpora are automatically segmented into spectrally stable units. The *second* task is the development of codebooks for the segments obtained in the first task. Assuming a source-system model for speech production, each segment is deconvolved into its source and system components. In linear predictive coding (LPC) based analysis, the source corresponds to the residual and system corresponds to that of the LPC coefficients. The system and source components of the segment are separately coded. To prepare the codebook of system parameters, the sequences of LPC vectors are clustered. The codebook for the source is prepared by quantizing the parameters such as pitch, gain, voicing etc. The *third* and final task in a segment vocoder is the encoding of the speech input. The input speech utterance is segmented into variable length segments. Each segment is quantized using the segment codebook. The index in the codebook that best matches the given segment is transmitted. The residual is encoded using any technique as in LPC10/MELP [4]. At the receiver, speech is synthesized using the sequence of LPC vectors taken from the codebook(based on the index received), and the decoded residual.

The two components of speech - source and system differ in their characteristics. Hence, the extent of compression that can be achieved for the two components is different. The source, which is fast varying is generally quantized frame by frame. Since the system characteristics are very slowly varying, quantization can be achieved over a set of frames or a segment. This facilitates better compression. In segment vocoders, the savings in bandwidth primarily comes from the choice of the segment; larger the segment, lower the bit-rate. The segment should not be too large that it cannot be modeled properly. The issue then is choosing an appropriate segmental unit which is reasonably stable and at the same time not too small. Generally, segments which are acoustically stable within the segment duration are chosen. For example, phonetic vocoders are built using phonemes [2] as basic segmental units. Attempts are also made to use diphones and multi-gram units as basic segments. In the research reported in this paper, we build a segment vocoder with a slightly larger unit, syllable. This requires that we be able to segment the speech signal into syllable-like units. The automatic segmentation methods like Maximum Likelihood(ML) segmentation and Spectral Transition Measure(STM) based segmentation are used to automatically segment a speech utterance into phoneme like units[5]. These methods assume piecewise stationarity of speech as an acoustic criterion for determining the segments. The criterion fails when there is a significant 'intra-segmental distortion.' The polyphonic nature of a syllable makes it difficult to use these segmentation methods. Further, the number of segment boundaries is fixed in the ML segmentation. The phoneme/syllable rate can vary quite significantly from one speech utterance to the other. Also, ML segmentation is an iterative method and is computationally intensive. In this paper, we use a signal processing technique called *Group Delay* (GD) based segmentation [6] which is successfully used to give syllable like units.

The segment codebook is prepared by clustering the segmental units obtained from a large speech corpus. Conventional methods include vector quantization (VQ) and its modifications [1][7][8][9][10] to cluster the segmental units in vector space. In VQ, the clusters are formed based on the spatial variation between the features of the segments but duration information is not modeled. Hence, duration and mapping analysis have to be addressed in the clustering process. When clustering using VQ, the sequence information is completely lost. Syllables being larger units, the sequence information is crucial. Therefore in this work, an acoustic clustering technique called unsupervised (Hidden Markov Model) HMM based clustering method [11] is used. In this clustering technique syllables which are acoustically similar fall into one cluster. Each cluster is identified by an HMM model. A codebook is generated based on the number of HMMs generated. The source(residual) is coded using MELP. The block diagram of the proposed segment vocoder is shown in Figure1.

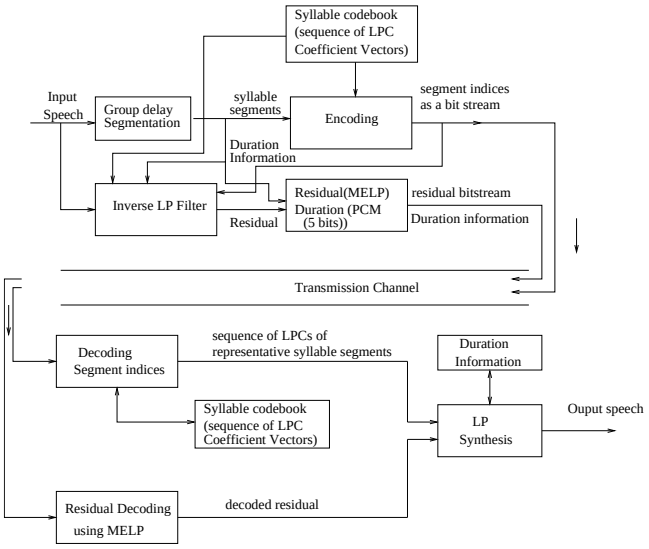

Figure 1: Block diagram of the proposed segment vocoder

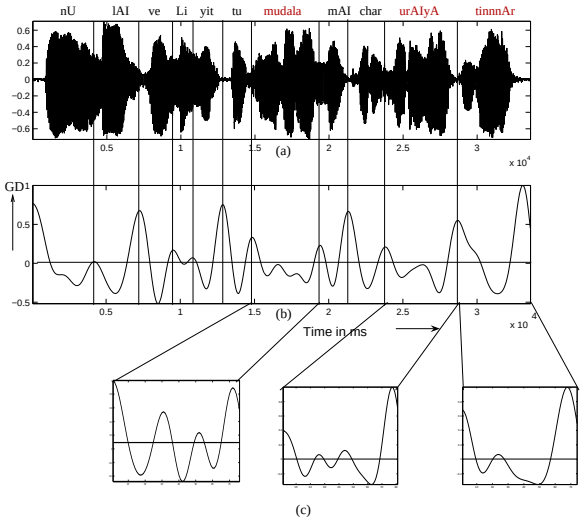

Figure 2: A two-level group delay (GD) based segmentation. (a) Speech signal (b) Segmentation - level one (c) Segmentation - level two

The organisation of the paper is as follows. In Section 2, we give a brief overview of group delay based segmentation. Section 3 details the segment codebook preparation. Here, we assume that the speech signal is deconvolved into source and system using LPC. The encoding and decoding processes are discussed as Experiments and Results in section 4. Finally, the conclusion and the issues to be addressed are discussed in section 5.

2. GROUP DELAY BASED SEGMENTATION

A signal processing method for segmenting speech into syllable like units uses the short term energy function (STE) of the speech signal. It is shown in [12] that if the signal is a minimum phase signal, the group delay computed on the spectrum will resolve the peaks and valleys well. The peaks and valleys in the group delay function of the signal will now correspond to the peaks and valleys in the short term energy function. If we consider a speech utterance as made up of syllables, the valleys in the energy spectrum correspond to the syllable boundaries. This is because a syllable has a high-energy voiced vowel region surrounded by low- energy unvoiced consonant regions. In other words, the valleys of the group delay function correspond to the syllable boundaries. To resolve the boundaries better, the inverted group delay is computed where the peaks mark the syllable boundaries. A number of research papers on segmentation of speech using the group delay based processing of STE have been reported in the literature [6, 13]. The most recent modification by [13], called the two-level group delay segmentation is used in this work. Two level segmentation gave an average performance of 87% for DDNEWS Tamil database.

The two-level group delay segmentation of a speech utterance is illustrated in Figure 2.

3. SEGMENT CODEBOOK PREPARATION

This process is carried out in three stages:

1. Clustering of syllable segments
2. Selection of representative syllable
3. Preparation of segment codebook entries

3.1 Clustering of syllable segments

The syllable segments obtained from a large speech corpus are automatically clustered using an unsupervised and incremental HMM based clustering algorithm proposed in [14]. After clustering is performed, syllable segments which are acoustically close to one another are grouped together into one cluster. Separate HMM model is built for each cluster.

The clustering is accomplished in two stages - initial cluster selection and unsupervised incremental training. A brief description of these two stages is given below:

3.1.1 Initial cluster selection

1. All syllable segments (say M) are used for clustering.
2. Features (13 MFCC/LPCC + 13 delta + 13 acceleration) are extracted using multiple frame-size and multiple frame-rate techniques[15] to compensate for the fact that there is only one example for every model.
3. An HMM model is initialized for each of the M syllables.
4. Viterbi decoding method is used to decode M syllables against M HMMs using 2-best criterion. This results in M pairs of syllable segments. Among these pairs, if a syllable segment is repeated in more than one pair, the other pairs are removed.
5. New models are created using these reduced number of pairs.
6. Steps 4 and 5 are repeated for m iterations (here 2) to get 2^m syllables in each cluster.

At the end of this stage, C_I initial clusters are formed where $C_I < M$. Initial cluster selection is implemented to ensure better and quick convergence of incremental training algorithm in the following stage.

3.1.2 Unsupervised incremental training

1. The model parameters of the initial C_I clusters are re-estimated using the Baum-Welch re-estimation.

2. All the M syllables are decoded against these re-estimated models using Viterbi decoding. Clustering is done based on the decoded sequence.
3. If a particular cluster is found to have less than a certain (here 6) number of examples, that cluster is removed.
4. Steps 1 and 2 are repeated until convergence is met. Convergence is reached when there is no migration of syllable segments from one cluster to another cluster as described in [16].

At the end of this stage, C clusters are formed where $C < C_I$. Each of the C clusters has a HMM.

The proposed clustering process is better demonstrated by using a colour palette diagram shown in Figure 3. Each of the M syllables is represented by a distinct colour in the colour palette. The colour model (analogy to HMM model) is built for each colour based on the features. Now, each colour is matched against the M colour models. The colours which have similar properties for example, the two shades of dark green are clustered into one group and the colour model of dark green is re-estimated using both the shades. Once the colours are clustered they are not used again for clustering in that iteration. This results in reduced number of clusters in the next iteration. This process of combining similar colours is carried out for some number of iterations till a set of distinct colours is obtained. This is represented as the palette of final colour models in the Figure 3. Here we can see that all shades of blue are modeled by a single blue colour model, all shades of green by a single green colour model and so on. Similarly, in the HMM based clustering process, similar syllables are clustered into one group and modeled by a single HMM model.

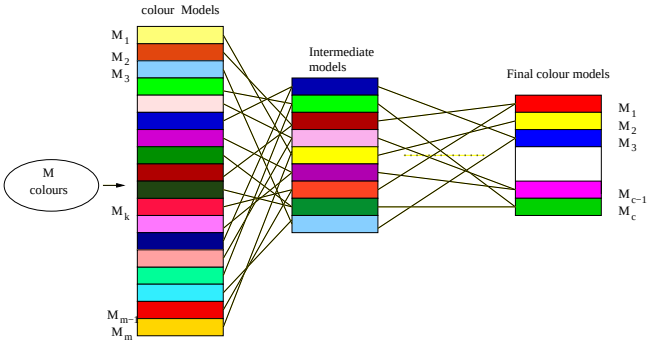

Figure 3: Colour palette analogy to explain the proposed HMM based clustering technique

3.2 Selecting representative syllable segment

To form the segment codebook, a 'representative' syllable is chosen for each cluster. This syllable represents the acoustic properties of all the syllable segments present in a cluster. The representative syllable is obtained as follows:

1. Let C be the number of clusters obtained, and let C_i denote the ith cluster and, N_i denote the number of syllable segments present in that cluster. s_{ij} denote the jth syllable segment in cluster C_i,
2. For each syllable s_{ij}, compute 10th order LPCC vectors (using 20ms frame-size and 10ms frame-shift).
3. The average Dynamic Time Warping (DTW) distance [17] distance D_{ij} for s_{ij} with each of the segments in the cluster is computed as follows:

$$D_{ij} = \frac{1}{N_i - 1} \sum_{k=1, k \neq j}^{N_i} DTW distance(s_{ij}, s_{ik})$$

4. The syllable segment which has minimum average DTW distance is chosen as the *representative syllable segment* for the cluster C_i.

The codebook of size C is prepared using the C representative syllables. The sequence of LPC coefficient vectors corresponding to the representative syllable segments are stored in the codebook. As the duration of syllables is not constant, the number of LPC vectors corresponding to each syllable are different. This codebook represents the system information. The residual is coded using standard MELP coding technique. This results in the increase of bit rate as the major portion of the bit-rate is occupied by residual bits. But, to check the sanity of the proposed techniques care is taken that residual is well modeled. Hence, MELP coding is chosen for residual coding.

4. EXPERIMENTS AND RESULTS

4.1 Encoding

The experimentation is carried on the Database for Indian Languages (DBIL) [18]. It consists of speech recordings in the Indian languages Telugu, Tamil and Hindi, sampled at 16 KHz and quantized at 16 bits/sample. In the present task, two Tamil databases are used - one is a single speaker (SS) database and the other is a multi-speaker (MS) database consisting of 8 speakers - 4 male and 4 female. The coder is implemented at 8 KHz and therefore the speech data is downsampled to 8 KHz before proceeding further. The segment codebooks of sizes 303 and 1583 respectively are obtained for the two databases.

When the input speech utterance to be coded is given, GD based segmentation is done to get the syllable segments. The segments are recognized using the HMM cluster models. The indices corresponding to the recognized cluster models are coded and transmitted along with the residual codebits obtained from MELP residual coding. The residual is coded for every 22.5 ms using 28 bits. As shown in the block diagram in Figure 1, the residual is obtained by passing the input speech signal through the LPCs of the corresponding representative syllables of the input syllable segments. This is done to ensure that the mismatch between the input syllable segments and representative syllables is captured in the error. This residual is coded where the pitch, jitter, voice and fourier magnitudes are computed and quantized using MELP algorithm [19]. This results in 1.2 Kbits for residual coding alone. If a syllable rate of 7 syllables/s is assumed and approximately 35 bits/s are allocated for duration information of the syllable segments, the overall bit-rate will be atmost 1.4 Kbps.

4.2 Decoding and Synthesis

Linear Prediction (LP) based synthesis is used to synthesize the speech at the receiver. The inputs to the synthesizer include the sequence of LPC vectors of the representative syllables corresponding to the transmitted sequence of segment codebook indices and the decoded residual using

MELP. Before synthesis, the durations of the representative syllables are to be matched with the corresponding syllable durations in the input speech utterance. The spectrograms of the original speech and the synthesized speech for a part of the test sentence are shown in Figure 4 and Figure 5 respectively. It is observed that the formant structure is preserved.

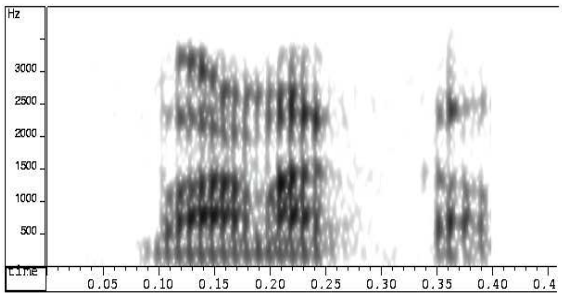

Figure 4: Spectrogram of the original speech utterance

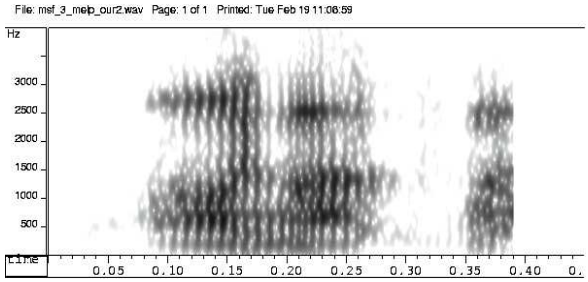

Figure 5: Spectrogram of the synthesized speech utterance

The comparison of performance results of the proposed coder and MELP codec at 2.4 Kbps is given in Table 1. The

Table 1: Comparison of PESQ scores for the proposed method and MELP 2.4 Kbps for the test utterances from Single Speaker(SS) and Multi-speaker(MS) databases

Test	Database Type	proposed method	MELP
1	SS	1.70	2.5
2	SS	1.45	2.5
3	MS	1.84	2.5
4	MS	1.70	2.4

lower PESQ scores for the proposed method is attributed to the mismatch of LPCs and the residual at the syllable boundaries. The informal listening tests showed good intelligibility. The synthesized sentences can be heard at the URL www.lantana.tenet.res.in/˜sadhana.

To improve the performance of the synthesized speech, the following method is used. During encoding, instead of selecting only the first best cluster for the syllable, 3-best clusters are selected. Representative syllables are chosen from the 3-best clusters. This leads to 3^N possible synthesized utterances, where N is the number of syllables in the utterance. Each of the 3^N sentences is synthesized using the decoding process described above. The syllable combination that yields the best PESQ [20] score is encoded and transmitted. Figure 6 illustrates the analysis-by-synthesis process for encoding a part of test utterance. The syllables corresponding to the path $- \times -$ are encoded and transmitted. The complexity can be significantly reduced by Viterbi based search technique with appropriate pruning.

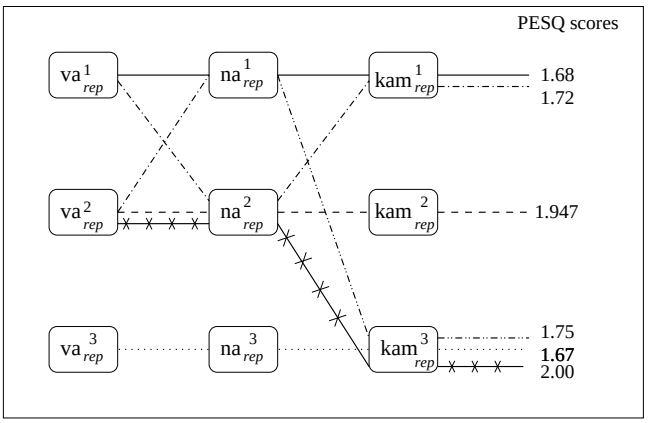

Figure 6: Different combinations of syllables for the utterance *"vanakkam"*

5. CONCLUSION

A novel approach to low bit rate segment vocoder is proposed. Syllable is chosen as the basic segmental unit. Group delay based segmentation is used for segmenting the speech. An unsupervised incremental HMM training algorithm is used to form the syllable clusters. The entries in the codebook correspond to the representative syllable units. There are number of issues that still remain to be resolved. We have observed that the choice of the representative syllable is not always the best. Further, the duration of a syllable is not well modeled. The HMMs generated during the clustering process could be used for synthesis.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Preeti Rao, Professor in Electrical Department, Indian Institute of Technnolgy Bombay for her insightful discussions which helped us understanding the problem at hand in more detail.

REFERENCES

[1] S. Roucos, R. M. Schwartz, and J. Makhoul, "A segment vocoder at 150 bits/s," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 61–64, 1983.

[2] J. Picone and G. Doddington, "A phonetic vocoder," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pp. 580–583, 1986.

[3] K. S. Lee and R. V. Cox, "A very low bit rate speech coder based on recognition/synthesis paradigm," *IEEE Transactions on Speech and Audio Processing*, vol. 9, July 2001.

[4] L.R.Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Eaglewood Cliffs, New Jersey, 1978.

[5] V. Ramasubramanian and T. V. Sreenivas, "Automatically derived units for segment vocoders," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 473–476, 2004.

[6] T. Nagarajan and Hema A Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," *EURASIP Journal of Applied Signal Processing*, vol. 17, pp. 2614–2625, 2004.

[7] S. Roucos, R. M. Schwartz, and J. Makhoul, "Vector quantization for very-low-rate coding of speech," *Proceedings of IEEE Globecom'82*, pp. 1074–1078, 1982.

[8] Y. Shikari and M. Honda, "LPC speech coding based on variable-length segment quantization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 36(9):1437–1444, 1988.

[9] Salim Roucos and Alexander M. Wilgus, "The waveform segment vocoder: A new approach for very low rate speech coding," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 10, pp. 236–239, 1985.

[10] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proceedings of IEEE*, vol. 73, pp. 1551–1583, 1985.

[11] T. Nagarajan and Hema A Murthy, "Language identification using parallel syllable-like unit recognition," *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 401–404, May 2004.

[12] Hema A. Murthy and B. Yegnanarayana, "Formant extraction from minimum phase group delay function," *Speech Communication*, vol. 10, pp. 209–221, 1991.

[13] A. Lakshmi and Hema A Murthy, "A syllable based continuous speech recognizer for Tamil," *Proceedings of INTERSPEECH-ICSLP*, pp. 1878–1881, Sept. 2006.

[14] G. L. Sarada, N. Hemalatha, T. Nagarajan, and Hema A Murthy, "Automatic transcription of continuous speech using unsupervised and incremental training," *Proceedings of INTERSPEECH*, pp. 405–408, 2004.

[15] G. L. Sarada, T. Nagarajan, and Hema A Murthy, "Multiple frame size and multiple frame rate feature extraction for speech recognition," *Proceedings of SPCOM*, pp. 592–595, 2004.

[16] T. Nagarajan and Hema A Murthy, "An approach to segmentation and labeling of continuous speech without bootstrapping," *Proceedings of National Conference of Communication*, pp. 508–512, 2004.

[17] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Pearson Education Pte. Ltd., Indian Branch, 482 F.I.E Patparganj, Delhi 110 092, India., 2003.

[18] "Database for Indian languages," *Speech and Vision Lab, IIT Madras, India*, 2001.

[19] Texas Instruments Inc, "2.4 kbps melp proposed federal standard speech coder, version 1.2," 1996.

[20] PESQ ITU-T P.862 (02/2001), "Perceptual Evaluation of Speech Quality (PESQ)," .