HIERARCHICAL SPEECH COMPRESSION FOR STORAGE - A TWO-LEVEL APPROACH

Peter Veprek and Alan B. Bradley
Department of Communication and Electronic Engineering
RMIT University, Melbourne, Australia

Abstract – In this paper, we investigate speech compression for storage. We propose a compression method operating on several levels of the speech signal hierarchy. The method is suitable for off-line compression of large closed corpora with known transcription. A realisation of the method using two levels of hierarchy is presented and evaluated. Results show that the method can provide high quality of reconstructed speech at low data rate for sufficiently large corpora.

INTRODUCTION

Structure of Speech

Speech signals, when recorded and digitised, can be analysed at various levels within an overall speech production hierarchy. First, there are samples of the signal itself. As such, individual samples are not of particular interest with the exception of waveform coding strategies (Deller et al., 1993).

The next level up in the hierarchy is the pitch epoch in the case of voiced speech and a frame in the case of unvoiced speech. Voiced speech is characterised by periodically vibrating vocal cords. This vibration modulates the flow of air passing through the vocal tract. The pitch period is the time between successive vibrations of the chords and the corresponding segment of the speech waveform is often referred to as a pitch epoch. Unvoiced speech can be segmented into (usually fixed-size) frames of similar duration. Due to the quasi-stationary nature of speech, each epoch can be described using the source-filter model by an excitation pulse and a set of formants or simply taken as a short waveform segment. It is clear that the epoch and frame repertoire of a single speaker is defined by the combination of different excitation pulses and vocal tract configurations and, although made up of a large set of possible states, it remains a limited set. Epochs and frames are often used as the basic units of speech production within speech synthesisers, coders, and voice converters (Hamon et al., 1989; Valbret et al., 1992; Veprek & Bradley, 1999).

Beyond the epoch and frame level analysis is the defined sequence of epochs and/or frames. While arbitrary sequences could be selected, diphones are commonly chosen as the unit at this level. By definition, each diphone represents the transition from one phoneme to another. The minimum number of diphones (in English) required for speech production is in the order of 1,700-1,800 and more if different acoustic and/or prosodic realisations of diphones are considered. The direct use of phonemes as a building block for speech production is an alternative, of course, with the benefit of having significantly fewer units (less than 100). However, due to strong coarticulation (particularly in English), phonemes are not adequate as a unit for speech generation and are consequently of little interest in the light of this research work.

Moving further up in the speech production hierarchy, a number of often overlapping, higher level units exist. To name a few, demisyllables and other sub-word units are frequently used in speech synthesisers (Pearson et al., 1998). Occasionally, entire words and/or phrases and their portions are considered as acoustic and/or prosodic units in some systems (Holm & Hata, 1998; Donovan et al., 1999; Lewis & Tatham, 1999; Stober, 1999). In addition, the speech signal can be examined from a number of different perspectives: acoustic, prosodic, linguistic, and so forth.

Traditional Speech Coding

Traditional speech coding techniques divide the digitised speech signal into short segments and encode these segments using diverse parameterisation techniques. The segment size varies from a single sample (in many waveform coders), through a fixed-size frame (in numerous linear-prediction based vocoders), to longer, often variable-length units such as phonemes (in several segment and phonetic vocoders). In most cases, a single type of segment is used although some techniques exploit redundancies at more than one level (e.g. adjacent sample and adjacent pitch cycle predictors utilised in some coders).

It is evident that typical speech coding techniques do not fully utilise the rich structural content of speech.

HIERARCHICAL COMPRESSION

Motivation and Expectations

Clearly, compression of speech using a single type of elemental unit is not an optimal approach because different types of redundancies exist at different levels within the hierarchical speech structure. For instance, at the sample level, redundancy is exhibited as high correlation between the amplitude of adjacent samples of voiced speech. Similarly, adjacent epochs and frames show relative resemblance in time and/or frequency domain, and different instances of a particular diphone have comparable acoustical properties.

Redundancies of diverse type exist at various levels within the speech hierarchy and require different models to represent and make use of them. Differential coding can be used at the sample level, while a long-term predictor may be used at the pitch epoch level, and stylisation of formant trajectories can be used at the diphone level.

A common technique extensively utilised for data compression is quantisation. During quantisation, similar instances of the data are grouped together and each group is collapsed into a single representative. Consider for instance pitch epochs/frames and diphones. Quantising the epochs/frames can significantly reduce the memory required for storage and the data bandwidth required for transmission of speech (Veprek & Bradley, 1999). However, compression of epochs/frames alone does not acknowledge the fact that the epochs/frames concatenate to form phonemes and diphones and that sequences of epochs/frames corresponding to the diphones are likely to occur repetitively - the existence of global patterns. Likewise, diphone compression alone can reduce the number of unique diphone realisations stored within a speech coder or synthesiser but ignores the fact that portions of different diphones are similar - the existence of finer granularity.

We propose a hierarchical method of speech compression where redundancies present in the speech signal at different levels are systematically exploited at those levels to achieve performance superior to other speech compression systems. The methodology is particularly suitable when compressing large volumes of speech for storage because of the higher compression ratio obtained for very large corpora and because of the coding delay introduced at all levels of the processing hierarchy.

Proposed Method Description - Two-Level System

When considering a speech compression method operating in a hierarchical manner on several levels of the speech structure, the choice of speech unit at each level and the number of levels in the system are critical decisions. The selected units must satisfy a number of requirements: the units must occur frequently, their instances must have similar characteristics, units should lend themselves to efficient compression, and so on.

Based on these requirements, we selected two specific units for a two-level hierarchical system. The first unit is a diphone and the second is the epoch/frame. Other candidate units include (demi)syllables, words, phrases, and others mentioned earlier. The number of unique diphones is not extremely high and a good match between adjacent diphones can be achieved by careful selection of their representatives and by smoothing diphone boundaries. This makes diphones suitable as units for concatenation in either the time or frequency domain. Epochs and frames are also appropriate units for speech production. Each epoch and frame can be described by a small set of (source-filter) parameters or taken as a short waveform segment. Epochs and frames are also well suited to prosodic modifications, either in frequency or in time domain.

As a compression methodology, we chose to use quantisation combined with prosodic modifications for both the diphones and epochs/frames. We denote this compression system as hierarchical diphone - epoch/frame compression using quantisation and prosody restoration (DEF-VQ).

Below is a block diagram of the proposed system (Figure 1). The top portion details the encoder while the bottom portion depicts the decoder. During encoding, compression is carried out in two stages. First, the speech signal is preprocessed to remove any offset and to normalise its amplitude. The speech is then aligned with the transcription of the signal and subsequently segmented into diphones. Following the segmentation, a codebook of diphone representatives is designed by training on the entire corpus and the input speech in quantised using the generated codebook. During quantisation, pitch, timing, and intensity profiles of the original diphones are extracted and stored in order to reconstruct the original prosodic characteristics during decoding. The diphone codebook obtained at

the first level is then the subject of epoch and frame compression at the second level. During the second level of processing, the speech signal within the diphones is classified into voiced and unvoiced portions. For the voiced portions of the diphones, pitch is determined and the speech signal is segmented into individual pitch epochs. Unvoiced portions are segmented into fixed-sized frames. Next, an epoch/frame codebook is designed and used to quantise the diphone codebook. Similar to the previous level, prosodic profiles are acquired and stored. More details about the epoch and frame compression technique can be found in (Veprek & Bradley, 1999). During decoding, speech is reconstructed in the reverse order. First, the diphone codebook is rebuilt by looking up the quantised epochs and frames, restoring their original prosodic characteristics, and concatenating them to form diphones. The speech is then reconstructed by retrieving and concatenating quantised diphones and modifying the overall prosody using the stored prosodic diphone profiles.

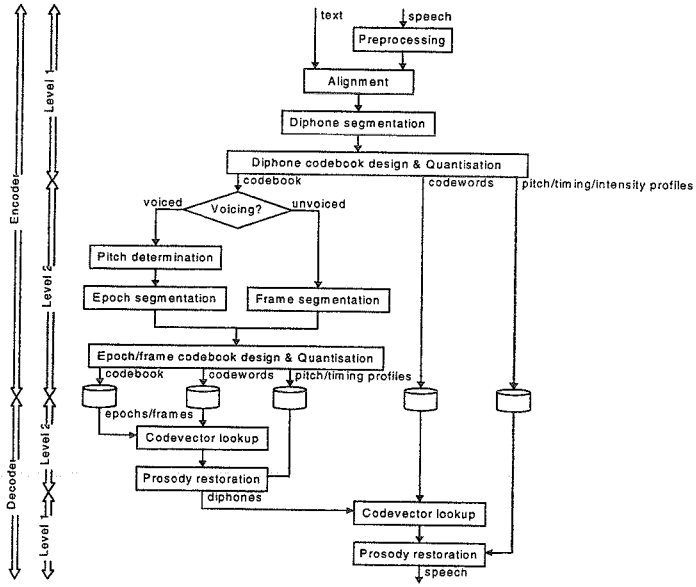

Figure 1. Overall block diagram of the proposed two-level system.

The prosodic modifications of epochs/frames performed during reconstruction of diphones and modifications of diphones during reconstruction of the speech consist of restoration of the unit (epoch/frame or diphone) duration, pitch profile of the unit, and overall RMS adjustment of the unit in case of the diphones. Timing and pitch are restored by storing the original values (individual pitch period and frame duration) in form of a profile and using them during reconstruction. Similarly, the RMS value of the original diphones is stored and used to scale amplitude of the reconstructed diphones.

Note that in this realisation, it was assumed that the speech transcription was available and used to guide the text-to-speech alignment that identified phonemes and diphones. Alternatively, existing speech recognition strategies could be used instead to generate the necessary boundaries and labelling. The text transcription can also be used when determining voicing and can be used to advantage as part of the distance measure used during quantisation.

EXPERIMENTAL EVALUATION

Data Set

The DEF-VQ method was evaluated using a speech database consisting of 211 unrelated sentences read spontaneously by a female speaker. The data amounted to 10 minutes 11 seconds of speech

and included 117,009 epochs and frames (68,658 epochs + 48,351 frames) and 5,896 diphones. There were 910 unique diphones present in the corpus. The speech was recorded at 11 kHz and sampled with 16-bit resolution.

Performance Measure

Performance of the DEF-VQ method was evaluated using two objective measures: the segmental signal-to-noise ratio (SNRseg) and the perceptual spectral distortion (SDp). Equations used to calculate the SNRseg were given in Veprek & Bradley (1999). The SDp was based on work of Hermansky and other researchers (Hermansky, 1990; Wang et al., 1991). SNRseg has been shown to correlate with subjective measures (Deller et al., 1993) and is typically used to measure the performance of waveform coders. Its use in our experimental evaluation is justified because the system operates in the time domain and is effectively attempting to reconstruct the original speech waveform. The perceptually motivated spectral distortion provides an objective measure that takes into account the following effects: non-linear auditory frequency warping, frequency-domain masking, equal-loudness, and perceived loudness.

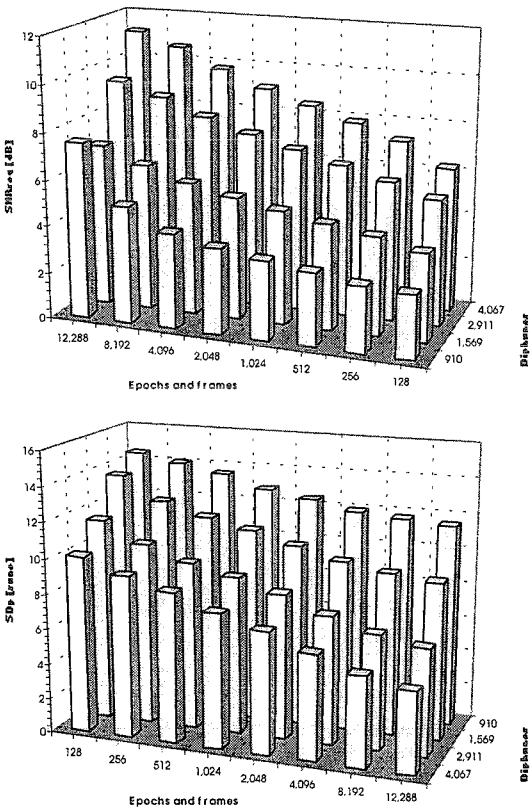

Figure 2. SNRseg and SDp as a function of the number of diphones and epochs/frames.

Results

Performance of the proposed system was evaluated as a function of the number of diphones utilised at the first level of quantisation and the number of epochs/frames used at the second level. The diphone codebook size ranged from 910 to 4,067 codevectors and the epoch/frame codebook size from 128 to 12,288 codevectors. Figure 2 below shows the SNRseg and SDp as a function of the number of diphones and epochs/frames.

Discussion

Generally, SNRseg increases monotonically with an increased size of the epoch/frame and/or diphone codebook. However, in one case, the SNRseg does not increase as intuitively expected. When the epoch/frame codebook size is fixed at 12,288 and size of the diphone codebook is increased from 910 to 1,569, the SNRseg drops down slightly. This can be attributed mainly to the fact that for the smaller diphone codebook there are more perfectly reconstructed sections of the output waveform coinciding with the analysis frames used during computation of the SNRseg thus affecting the measure. Informal listening tests showed that the perceptual spectral distortion of less than 8 sones corresponds to good quality of reconstructed speech.

Projection for Larger Corpora

Following the evaluation, we calculated the data storage requirement for this compression strategy. The data rate consists of that necessary to encode the epoch/frame codebook, sequence of epoch/frame codewords, prosodic profiles of epochs/frames, sequence of diphone codewords, and prosodic profiles of diphones. For the evaluation corpus used, the overall data rate ranges from less than 1 kbps for small diphone and epoch/frame codebooks to about 12 kbps for large codebooks.

Next, we were interested to see how the data storage requirement depends on the corpus size. From earlier investigations we knew of the trend of better speech-quality-to-storage-requirement ratio being achievable for larger corpora (Veprek & Bradley, 1999). In view of this we subdivided the data storage requirement into five contributing categories: diphone codeword storage, diphone profiles, epoch/frame codebook, epoch/frame codewords, and epoch/frame profiles; and we formulated expressions to calculate each storage need as a function of several parameters. We then computed the individual data rates for an increasing corpus size for the scenario in which 4,067 diphones and 12,288 epochs and frames were used. Results of this computation are graphically illustrated in Figure 3 below.

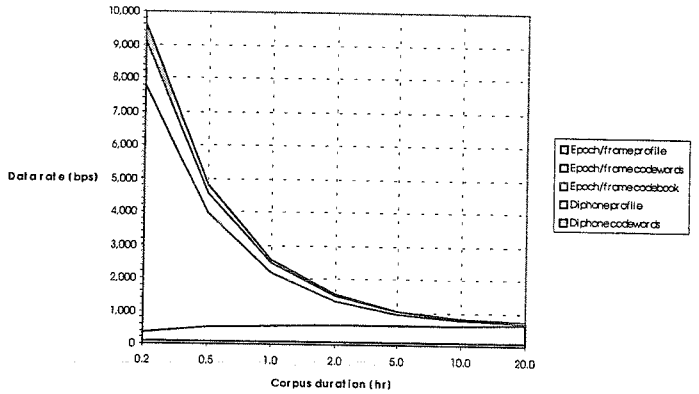

Figure 3. Data rate as a function of the corpus duration (D 4,067 & EF 12,288)

The obtained results are truly accurate only for the particular speech corpus and the described compression conditions. Nevertheless, although other corpora and conditions will undoubtedly yield different data rate values, the overall trend will remain the same. We can, therefore, make several general conclusions that will hold for the investigated as well as other cases. Two specific points are worthy of noting. First, in this case, we are examining a hierarchical system consisting of only two levels - diphone and epoch/frame. Second, the system lacks any true entropy coding. In the storage rate calculation we included a 10% saving due to efficient bit allocation but obviously better results can be achieved using more sophisticated coding techniques. Both of these facts mean that even better results can be expected from a system including more compression levels and entropy coding.

The first general conclusion we can make is that it is clear that for corpora larger than about 5 hours the effective data storage rate drops down to approximately (actually below) 1 kbps. Even if a larger number of representative diphones and a larger epoch/frame codebook were used, the total data storage rate would still be within the 1 kbps neighbourhood. We can see that even if the number of diphones in the first level and the number of epochs/frames in the second level were doubled, the increase in the data rate would still be rather small (around 160 bps for the 10-hour corpus and 80 bps for the 20-hour corpus) resulting in the overall data rate well under 1 kbps for 10-hour and larger corpora.

Second, the data storage break-up into the five specific components indicates that for smaller corpora (less that 2 hours) most of the data storage requirement is consumed in storing the epoch/frame

codebook while for larger corpora (over 5 hours) the cost of encoding the diphone profiles becomes dominant. Based on the observation of the data break-up tendency for larger corpora, we can conclude that for very large corpora, special attention must be paid to compression of the diphone profiles because they consume the dominant portion of the total data rate.

CONCLUSIONS

Hierarchical speech compression is a concept in which compression is carried out in several stages at different levels of the speech signal structure. While the presented system consisting of only two levels (epoch/frame and diphone) is merely one possibility for a hierarchical arrangement, it is shown to have practical use for speech storage applications because of significant redundancies amongst units at those levels and due to the high frequency of occurrence of such units in speech. Data models exploiting the redundancies of speech units in order to achieve compression can vary and the quantisation utilised in the presented system is only one of a number of possible implementations. Also, a hierarchical compression system does not have to operate exclusively in the time domain and it can employ both time and/or frequency domain techniques as appropriate.

The proposed system provides a powerful framework for hierarchical speech compression and is particularly suitable for processing large speech corpora for efficient storage. Based on the evaluation and analysis we can assert that high speech quality at low bit rates can be obtained for corpora of approximately 5 hours and more. For smaller corpora, greater data rates are required.

To generate the best set of representative units, epochs and frames for example, a large number of computations is generally required. The number increases with an increase of both the corpus and codebook size. In the hierarchical arrangement however, the optimum codebook for any given level within the speech analysis hierarchy is computed using just the entire codebook of the next higher level within the hierarchy. In the examined case, the epoch/frame codebook needs to be generated not from the entire corpus but only from the diphone codebook obtained previously. This would also be true in case of diphones if another (prior) level was present in the system. As a result, the hierarchical architecture leads to reduction in computation complexity at lower levels.

When examining break-up of the total data storage requirement, it is evident that for increasing corpus size, the bit rate consumed by the diphone profiles becomes dominant. Consequently, special consideration must be given to the diphone profile compression. Finally, the system could be augmented, for instance, by addition of a sample level. The sample level could employ techniques such as ADPCM to further reduce the overall bit rate requirement.

ACKNOWLEDGEMENT

The speech database used for experimental evaluation presented in this paper was obtained from Mr. S. Pearson and Dr. N. Kibre from the Panasonic Speech Technology Laboratory of Panasonic Technologies, Inc. located in Santa Barbara, CA, USA.

REFERENCES

Deller, J. R., Proakis, J. G. & Hansen, J. H. L. (1993) Discrete-time processing of speech signals.
Donovan, R. E., Franz, M., Sorensen, J. S. & Roukos, S. (1999) "Phrase splicing and variable substitution using the IBM trainable speech synthesis system", ICASSP.
Hamon, C., Moulines, E. & Charpentier, F. (1989) "A diphone synthesis system based on time-domain prosodic modifications of speech", ICASSP.
Hermansky, H. (1990) "Perceptual linear predictive (PLP) analysis of speech", JASA.
Holm, F. & Hata, K. (1998) "Common patterns in word level prosody", ICSLP.
Lewis, E. & Tatham, M. (1999) "Word and syllable concatenation in text-to-speech synthesis", EuroSpeech.
Pearson, S., Kibre, N. & Niedzielski, N. (1998) "A synthesis method based on concatenation of demisyllables and a residual excited vocal tract model", ICSLP.
Stober, K., Portele, T., Wagner, P. & Hess, W. (1999) "Synthesis by word concatenation", EuroSpeech.
Valbret, H., Moulines, E. & Tubach, J. P. (1992) "Voice transformation using PSOLA technique" Speech Communication, vol. 11, nos. 2-3, pp. 175-187.
Veprek, P. & Bradley, A. B. (1999) "Speech compression by vector quantization of epochs", ISSPA.
Wang, S., Sekey, A., & Gersho, A. (1991) "Auditory distortion measure for speech coding", ICASSP.