

COMPRESSION OF SPEECH FOR MASS STORAGE USING SPEECH RECOGNITION AND TEXT-TO-SPEECH SYNTHESIS

John Dines, Sridha Sridharan and Miles Moody
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
j.dines@qut.edu.au s.sridharan@qut.edu.au
m.moody@qut.edu.au

ABSTRACT: In this paper a speech compression algorithm is presented that utilises speech recognition and text-to-speech synthesis technology to code speech at very low bit rates suitable for mass storage applications. The system relies on a word level transcription and carries out a phonetic alignment of the signal using a lexicon of pronunciations. Speech is synthesised at the decoder by concatenating diphone segments from a speaker dependent database. Prosody and energy information is extracted from the original source speech and compressed using a low rate scheme. In order to synthesise speech that is perceived as being produced by the target speaker it is necessary that a speaker transformation scheme be adopted. Two speaker transformation schemes have been tested: a vector quantisation scheme and a direct estimation mapping scheme. Rates of 220 to 400 bps have been achieved using this approach. Informal subjective testing was carried out to compare the synthesised speech of several coding schemes in terms of intelligibility and speaker recognisability.

1. INTRODUCTION

The application of speech recognition and text-to-speech (TTS) technology to very low bit-rate compression of speech signals has been recognised as one of the most promising avenues of investigation for producing natural and intelligible speech below 500 bps (Lee and Cox, 1999; Ribeiro and Trancoso, 1997; Tokuda et al., 1998).

This class of segmental vocoders are known as phonetic vocoders. Compression is achieved by segmenting the input speech into its phonetic transcription at the encoder and decompression by synthesising the phonetic units from the transcription at the decoder. Most of the computational load of the system is at the encoder carrying out speech recognition, while the decoder is significantly less complex as phonetic units are synthesised from a database of speech units.

Several methods have been used for the synthesis of speech units at the decoder including synthesis from HMM statistics (Tokuda et al., 1998), from a codebook of normalised phonetic units (Ribeiro and Trancoso, 1997) and using a Harmonic plus Noise Model TTS system (Lee and Cox, 1999). In our work we have based the decoder on a pulse excited LPC text-to-speech synthesis system for the synthesis of the short term spectrum. Signal energy and prosodic information are extracted directly from the speech signal.

This paper is organised as follows: Section 2 describes the compression and decompression algorithm, Section 3 details the results in which we compare the performance of the system under various configurations and bit-rates, and finally in Section 4 we make some general concluding remarks and discuss possible future improvements to the vocoder.

2. COMPRESSION AND DECOMPRESSION ALGORITHMS

The compression/decompression system is depicted in Figure 1. The coder carries out two primary functions: segmentation of the speech into its phonetic transcription and pitch and voicing analysis. The former of these produces a sequence of phoneme labels and their durations which enable reconstruction of the short term speech spectrum. The latter enables reconstruction of the excitation signal. In addition,

energy is extracted, as it has been observed that accurate preservation of the signal amplitude plays a crucial role in maintaining naturalness and intelligibility of the reconstructed speech. Voice conversion is carried out in order to synthesise speech that sounds like that produced by the target speaker.

The decoder comprises a text-to-speech engine, which carries out synthesis of the speech by concatenation of time-warped diphone segments of speech. The excitation signal is synthesised using a simple binary excitation model. The amplitude of the synthesised speech is modified according to the stored gain information.

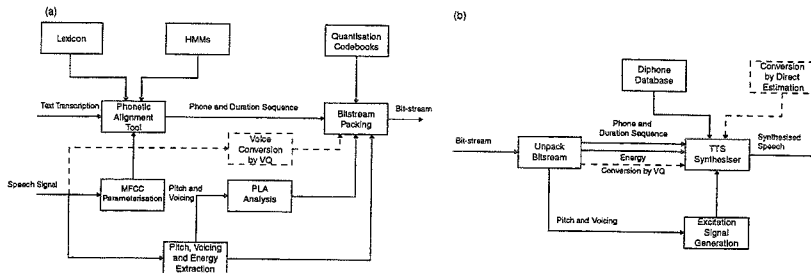


Figure 1: The compression and decompression system. (a) Encoder (b) Decoder

2.1. Speech Transcription

For speech coding applications, a major problem with the use of diphone concatenation is that phonetic errors generated by the speech recogniser result in poor spectral matching between the original and synthesised spectra, especially when the desired diphone entry has not been stored in the TTS database. For this reason it has been assumed that a text transcription of the speech is available. This assumption is realistic in several situations — for example where speech is read from a manuscript or where the speech has been transcribed previously by an operator. A lexicon of word pronunciations can be used to produce the phonetic transcription, which may be accurately aligned with the speech using a Hidden Markov Model (HMM) speech recognition engine.

Segmentation of the speech data is carried out by a HMM speech recogniser. The recogniser is based upon a set of 42 left-right monophone models (including silence and short pause models) using a 10 coefficient MFCC parameterisation (plus energy, delta and delta-delta terms, giving a total of 33 coefficients). These models were initialised using the male speakers of the TIMIT phonetically aligned corpus, then triphone context dependent models were trained using 9 long term male speakers from the Continuous Speech Recognition Corpus (WSJ1). The states in these models were tied using the decision tree technique to result in approximately 5000 distributions and 1200 unique context dependent models.

Phonetic alignment of the word level transcription requires a lexicon of pronunciations be available. The *CMUDICT 4.0 American English Lexicon* (1996,1997) was used to convert word to phonetic level transcriptions. The dictionary comprises approximately 100,000 entries resulting in approximately 20,000 possible triphone models which can be synthesised from the trained set of models using the decision tree.

To create accurate alignments of the phonetic transcription a forced alignment was carried out. Inspection of the automatically aligned speech showed that in most cases the phonetic boundaries were located within a few frames of the hand labeled boundary by the alignment tool. An example is shown in Figure 2.

2.2. Pitch, Voicing and Energy Parameters

Pitch tracking is carried out using the super resolution pitch detection algorithm (Medan et al., 1991). Following post-processing of the raw pitch values, the pitch contour is stylised using the Piecewise Linear

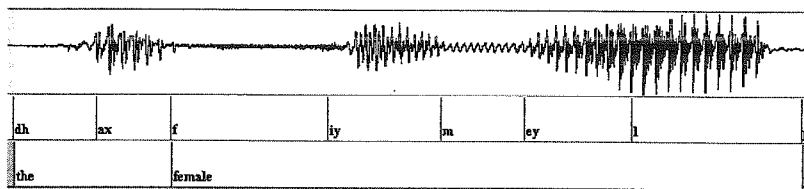


Figure 2: Word and phoneme level alignments of a segment of speech.

Approximation Method (Lee and Cox, 1999). This method uses linear segments to represent the pitch contour with a maximum allowable error between the approximated and actual contour.

Energy information is encoded once every four frames (20 ms). This appears to give a relatively good reconstruction of the gain contour while maintaining a low bit rate. The quantisation scheme assumed to code this information is detailed in Section 2.4.

2.3. Speaker Adaptation

Two speaker adaptation schemes have been implemented that attempt to map the speech produced by the text-to-speech system to that of the recorded speech that is being compressed. The first of these techniques is a vector quantisation method that simply adds a transformation vector to the synthesised spectral envelope. The second technique uses a Joint Density Gaussian Mixture Modeling technique (Kain and Macon, 1998b). This algorithm has been previously shown to effectively perform a transformation between two speaker spaces for text-to-speech synthesisers.

2.3.1. Voice Conversion by Vector Quantisation

This method of voice conversion uses a transformation vector which is added to the synthesised spectral envelope. The transformation vector is calculated once per phoneme in the transcription, and is simply the difference between the weighted means of the target and source parameters for the phonetic instance. The purpose of the weighted mean is to emphasise the values at the centre of the phoneme while placing less emphasis on the values at the beginning and end. This is because the centre of the phonemes are areas of spectral stability where we would expect the most speaker dependent characteristics to be present. The weighting function is shown in Eqs. (1)-(3).

$$k(t) = \left| \frac{2t - t_2 - t_1}{t_2 - t_1} \right|, \quad t_1 \leq t \leq t_2 \quad (1)$$

$$w_{k(t)} = \exp(-c k(t)), \quad 0 \leq k(t) \leq 1 \quad (2)$$

$$\bar{x}_w(t) = \frac{\sum_{t=t_1}^{t_2} w_{k(t)} \Delta C(t)}{\sum_{t=t_1}^{t_2} w_{k(t)}} \quad (3)$$

where t_1 and t_2 are the time indices of the beginning and ending of the phone. $k(t)$ rescales t such that 0 lies at the centre of the phone and 1 lies at the phone boundaries. c is an arbitrary constant (1 in our experiments) and $\Delta C(t)$ is the difference between source and target spectral parameters at time t . $w_{k(t)}$ is our weighting function and $\bar{x}_w(t)$ is the transformation vector (a weighted mean).

Likewise, at the synthesiser, speech is modified by adding the transformation vector to the synthesised spectral parameters using the same weighting function:

$$C_m(t) = C_s(t) + w_{k(t)} \bar{x}_w(t) \quad (4)$$

where $C_s(t)$ is the synthesised spectral parameter and $C_m(t)$ is the mapped spectral parameter. $\bar{x}_w(t)$ is the quantised transformation vector.

2.3.2. Voice Conversion by Direct Estimation

In our work the similarities between phonetic and text-to-speech (TTS) vocoders has been investigated to incorporate an adaptation algorithm based upon text-to-speech voice transformation. This scheme requires that the speaker adaptation information be stored only once in the form of a joint density Gaussian Mixture Model.

The papers by Kain and Macon (1998a,b) show that a direct estimation method that uses locally linear mappings between source and target speakers can give good results for spectral conversion in TTS systems. The speaker transformation function is based upon a Joint Density Gaussian Mixture Model, Eq. (5), trained on source and target parameter sets, $z = [x_t^T y_t^T]^T$. The conversion function is given by (6) and is simply the expectation of y given x .

$$p(x_t) = \sum_{i=1}^m \alpha_i N(x_t; \mu_i, \Sigma_i) \quad (5)$$

$$F(x_t) = E[y|x] = \sum_{i=1}^m P(C_i|x_t) [\mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x_t - \mu_i^x)] \quad (6)$$

where $N(x_t; \mu_i, \Sigma_i)$ denotes a Gaussian distribution with mean vector, μ_i , and covariance matrix, Σ_i , and α_i is the mixture weight. $P(C_i|x_t)$ is the conditional probability that the feature vector, x_t , belongs to the class C_i . μ_i^y and μ_i^x are the means of the target and source distributions respectively for class C_i and Σ_i^{yx} and Σ_i^{xx} are the cross- and auto-covariance matrices for class C_i .

2.4. Quantisation of the Bit-stream

Several configurations of the speech coder have been evaluated, these have been summarised in Table 1. These schemes vary in their quality and bit-rate, ranging from Scheme 1 which simply attempts to code intelligible speech at the lowest rate by including only the phone sequence and prosodic information, while Schemes 3 and 4 attempt to reproduce speech that is both intelligible and sounds like it was produced by the original speaker by also including energy and speaker adaptation information.

The phoneme index may take one of 42 possible values, at an average phoneme rate of 10 per second resulting in an average allocation of 55 bps using Huffman coding. The phoneme duration is scalar quantised with 5 bits (a separate codebook for silence durations) giving an average bit rate of 50 bps.

Coding of pitch and voicing information using piecewise linear approximation stores a (scalar quantised) pair of values: the linear segment duration and the starting pitch value for that segment. At a voiced-to-unvoiced transition, the final pitch value for the preceding segment is transmitted, plus a duration of zero. Sufficient prosody information is transmitted with each phoneme such that the excitation signal for that segment may be completely reconstructed. This results in a measured bandwidth of 115 bps for pitch and voicing information given an allocation of 5 bits for the pitch and 5 bits for the duration values.

The energy values (produced once every 20ms) are concatenated into vectors of five and quantised using a 10 bit VQ codebook. This gives an average rate of 100 bps.

Voice conversion using vector quantisation requires that a 12 bit index be stored once for each phoneme occurrence (except for silences), which results in a bit rate of 80 bps for the measured phoneme rate. The voice conversion using direct estimation need only be stored once at the beginning of the compressed speech, hence it has not figured in the bit-rate calculations.

The bit allocation scheme is summarised in Table 2.

2.5. Synthesis Using Text-to-Speech

Synthesis is carried out using a pulse excited LPC text-to-speech synthesiser, based upon the Festival Speech Synthesis System (Black et al., 1999) and OGIresLPC synthesiser plug-in (Macon et al., 1997). The synthesiser has been modified to allow artificial generation of the excitation signal using a binary excitation model which gives greater control over pitch and voicing modification than the original residual

Scheme	Phone, Duration & Pitch	Energy	Voice Conversion	
			Direct Est.	VQ
1	X			
2	X	X		
3	X	X	X	
4	X	X		X

Table 1: The coding schemes that were tested.

Parameter	Avg rate
Phoneme	55
State Duration	50
Energy	100
Pitch & Voicing	115
VQ Voice Conversion	80
Scheme: 1	220
2	320
3	320
4	400

Table 2: Rates for the proposed coder configurations (bps).

excitation technique. Furthermore, the database of speech segments is greatly reduced in size as there is no longer a requirement to store the residual signal speech data. A disadvantage of this method is that synthesised speech will tend to sound more artificial, although this could later be rectified by using a mixed excitation scheme.

The short term speech spectrum is constructed by concatenating diphone segments which are time warped to produce phone durations matching those generated by the speech alignment tool. The synthesiser performs post processing of the concatenated segments to reduce discontinuities at the joins.

3. EXPERIMENTAL RESULTS

Informal subjective evaluation of the coding schemes was carried out to rate the coding schemes in terms of intelligibility and speaker recognisability. The coders were rated on a five-point Mean Opinion Score (MOS) scale. Five untrained listeners were first given several examples in order to give them an indication of best and worst cases to expect, and then were asked to rate ten sentences produced by each coding scheme to assess intelligibility and three sets of 6 sentences (the original speech, the benchmark and the four coding schemes) to assess speaker recognisability results. An unquantised LPC based vocoder was used as the benchmark with which results of the 4 coding schemes were compared. The results of the experiments are shown in Figures 3 and 4.

It can be seen that all of the coding schemes perform with little degradation to the intelligibility compared with the benchmark coder, although all schemes rated relatively low compared to the original speech (a perfect score of 5). This may be attributed to the fact that the speakers were all unused to listening to synthesised speech. Improving the quality of synthesised speech by using a mixed excitation scheme would be expected to improve these results.

Speaker recognisability results produced the expected outcome, with better subjective test results for schemes using voice conversion techniques. Furthermore it can be seen that including gain information alone resulted in an improvement in both intelligibility and speaker recognisability results.

4. CONCLUSIONS AND FUTURE WORK

In this paper we have shown that speech may be compressed at very low bit-rates in the range of 220-400 bps using speech recognition and text-to-speech synthesis. The resulting speech is intelligible and may be perceived as being produced by the original speaker when speaker transformation techniques are utilised. Prosodic characteristics of the speech are reconstructed naturally although the simple excitation model results in somewhat artificially sounding synthesis.

There are several avenues available to improve the overall performance of this system. Better quality synthesis could be obtained by introducing a mixed excitation scheme with an acceptable increase in bandwidth of the coder. Research is also being carried out to develop more effective speaker transformation schemes. Further advancements in text-to-speech synthesis technology will no doubt also provide a platform for producing better quality speech compression using this research.

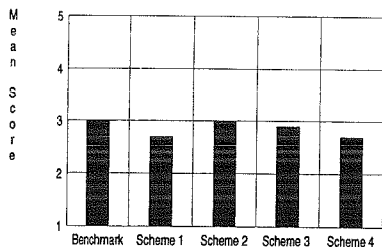


Figure 3: Intelligibility results.

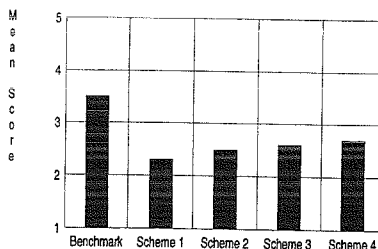


Figure 4: Speaker recognisability results.

5. ACKNOWLEDGEMENTS

Experiments were carried out using a TD-PSOLA synthesiser based upon the Festival Text-To-Speech Synthesis System (Black et al., 1999) and OGiresLPC plug-in residual excited LPC synthesiser (Macon et al., 1997). This work was supported by the CSIRO Division of Telecommunications and Industrial Physics.

6. REFERENCES

- Black, A. W., Taylor, P. and Caley, R. (1999), *The Festival Speech Synthesis System*, Centre for Speech Technology Research, University of Edinburgh, UK.
- CMUDICT 4.0 American English Lexicon (1996,1997), Centre for Speech Technology Research, University of Edinburgh, UK.
- Kain, A. and Macon, M. W. (1998a), Personalising a speech synthesiser by voice adaptation, *Proc. ESCA/COCOSDA International Speech Synthesis Workshop*.
- Kain, A. and Macon, M. W. (1998b), Spectral voice conversion for text-to-speech synthesis, *Proc. ICASSP-98*.
- Lee, K. and Cox, V. (1999), TTS based very low bit rate speech coder, *Proc. ICASSP-99*.
- Macon, M., Cronk, A., Wouters, J. and Kain, A. (1997), OGiresLPC: Diphone synthesiser using residual excited linear prediction, *Technical report*, CSLU, Oregon Graduate Institute, Portland, OR.
- Medan, Y., Yair, E. and Chanzan, D. (1991), Super resolution pitch determination of speech signals, *IEEE Trans. Signal Processing*, Vol. 39, pp. 40-48.
- Ribeiro, C. M. and Trancoso, I. M. (1997), Phonetic vocoding with speaker adaptation, *Proc. of EUROSPEECH-97*, pp. 1291 - 1294.
- Tokuda, K., Masuko, T., Hiro, J., Kobayashi, T. and Kitamura, T. (1998), A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques, *Proc. ICASSP-98*, pp. 609 - 612.