BETWEEN RECOGNITION AND SYNTHESIS - 300 BITS/SECOND SPEECH CODING

M Ismail and K Ponting
DERA Malvern
Speech Research Unit
St. Andrews Road, Malvern, Worcs, WR14 3PS, England, U.K.
Email: ismail@signal.dra.hmg.gb

ABSTRACT

This paper describes a system for speech coding designed to operate at 300 bits/sec and below. A continuous speech recogniser is used to transcribe incoming speech as a sequence of sub-word units termed acoustic segments. Prosodic information is combined with segment identity to form a serial data stream suitable for transmission. A rule-based system maps segment identity and prosodic information to parameters suitable for driving a parallel formant speech synthesiser. Acoustic segment Hidden Markov Models (HMMs) are shown to perform as well as conventional phone HMMs during recognition. A segment error rate of 3.8 % was achieved in a speaker-dependent, task-dependent configuration. An average data rate of 262 bits/sec was obtained. Speech from the synthesiser was better than obtainable from a purely textual representation though not as good as 2400 bit/sec Linear Predictive Coding (LPC) vocoded speech.

INTRODUCTION

LPC based vocoders have been used successfully for a number of years providing good quality speech over a range of data rates. For very low data rates (sub 800 bits/sec) LPC coded speech is very poor due to the reduced number of coefficients available for accurate modelling of the spectrum. Information coding theory tells us that the human-human speech communication data rate is of the order of 40-80 bits/sec. Clearly, there is a discrepancy between what is possible currently and what is possible theoretically.

Although LPC coding models the human vocal tract, no constraints are imposed on the possible configurations that the model can represent. Humanly impossible tract shapes are equally as valid as real tract shapes. Some LPC based coders impose a smoothing function on the LPC parameters exploiting inter-frame correlation due to the smooth manner in which the vocal tract changes shape. However, even this does not limit the modelling to sounds observed in a given language, so a sequence of meaningless sounds can still be modelled. This inherent redundancy increases the entropy, and hence the number of bits required for encoding, of the system. Attempts have been made at achieving very low data rates based on a more sophisticated modelling scheme [1][2].

A different approach is to develop a rich but compact inventory of sounds occurring in a given language. These basic units can then be used to describe any utterance in that language. The accuracy of the description, e.g. emotional state, depends on the richness of the model. Since the number of sounds produced in a language is far fewer than all the sounds that can be produced by the human vocal system encoding of speech using such a representation should be more efficient than LPC coding.

In this paper we present a speech coding system based on automatic speech recognition of the speech signal. The current system uses speaker dependent, task dependent speech models to transcribe incoming speech as a sequence of sub-word units. Prosodic information is combined with the transcription to form a sub 300 bits/sec data stream. At the receiver, a speech synthesiser is used to reconstruct the original message.

Our motivation for pursuing speech coding at sub 300 bits/sec is based on the observation that under adverse communication channel conditions the bandwidth available for speech transmission may be greatly curtailed. In extreme conditions no communication may be possible at all, but in a HF communications channel up to 300 bits/sec may be sustainable. Presently, there appears to be a lack of real-time speech coding systems operating satisfactorily in the region of 200-600 bits/sec.

A SEGMENT VOCODER

A continuous speech recognition system is used to transcribe incoming speech as a sequence of sub-word units termed acoustic segments. In parallel, prosodic information is extracted for subsequent use during speech synthesis. Speech synthesis is performed by a parallel formant synthesiser in tandem with a rule-based

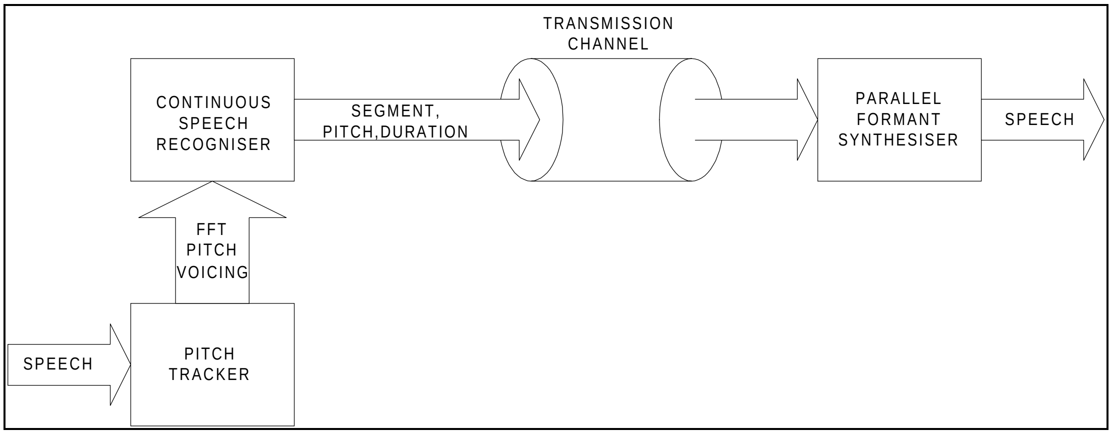

Figure 1: Segment Vocoder

system mapping acoustic segments to formant parameters. Figure 1 gives an overview of the segment vocoder.

This system uses a hybrid of two fairly well understood techniques, HMM recognition and synthesis-by-rule (SbR). Soong [3] uses HMM based recognition and uses the recognition training material to derive synthesis parameters. Other attempts have used recognition-by-synthesis-by-rule (RbSbR) [4] in which, rather than matching incoming speech against hypotheses corresponding to sequences of HMM states, it is matched against hypothetical speech patterns generated by the SbR system. Our system uses a HMM based continuous recogniser and a parallel formant synthesiser for speech synthesis.

This approach was taken for two reasons, HMM based synthesis [5] requires considerable training of the synthesiser to a particular talker, in [5] this required 34 hours of CPU time although more recently speaker adaption of HMM-based synthesis has been demonstrated from a few sample sentences [6]. RbSbR is computationally very expensive and not readily available as a real-time system whereas a continuous HMM based system can run on a moderately powerful desktop computer.

Speech Recognition

Acoustic segments are an extension to the familiar SAMPA phoneme notation allowing for finer modelling of speech sounds. Unvoiced and voiced plosives are expanded into a closure, burst, aspiration and closure, burst representation respectively. Each expansion is subject to context and a set of rules define expansion of a phoneme to its corresponding segment(s) subject to context. Holmes, Mattingly and Shearme[7] define seventy-nine acoustic segments and describe a rule-based approach for deriving formant parameters, suitable for driving a parallel formant synthesiser. In our recognition system a sub-set of sixty-four acoustic segments are used with diphthongs being represented by a single model rather than two models as in the HMS[7] system.

Presently the recogniser is used in a task dependent, speaker dependent configuration. This is to ensure a sufficiently high segment recognition rate for intelligible speech synthesis. Zue[2] suggests a phoneme recognition accuracy of 85% is required for intelligible speech synthesis from a phoneme sequence. Recognition is performed using segment level Hidden Markov Models (HMMs). A dictionary defines all possible words in the task in terms of their segment spelling. An airborne reconnaissance task (ARM) [8] with a five hundred word vocabulary is used in the present system.

An alternative approach for small fixed vocabularies would be to use whole word recognition and synthesis. However, for our longer term goal of task (vocabulary) independence or rapidly changing vocabularies this system has severe limitations.

Defining words in terms of sub-word units allows for addition of new words to the dictionary without having to train new models. Whole word concatenation for synthesis is only feasible for small vocabularies and suffers from a similar problem to whole word recognition when new words are added, namely collection of additional data.

Conventional phoneme HMMs are three state left to right no skip models. At a frame rate of 100 frames/sec a minimum duration of 30 ms (1 frame/state) is needed per model. In the segment modelling scheme an unvoiced plosive is represented as three separate models, closure, burst and aspiration. Using the standard 3 states per model would result in an unvoiced plosive having a minimum duration of 90 ms when using acoustic segments. A different state allocation strategy was investigated, allocating a single state to each of the three segments of a plosive.

Pitch Analysis

Although the synthesiser is capable of generating speech from a segment representation the naturalness is greatly impaired since no prosodic information is available from the original speech signal. A cepstrum based pitch tracker is used to provide pitch and voicing information to the synthesiser.

Cepstrum analysis and its application to pitch estimation is described in the paper by Noll[9]. The cepstrum is calculated by taking the Fourier transform of the of the logarithm of the spectrum. The feature vectors used in the recognition process are also derived from the same log spectral representation.

Speech Synthesis

Parallel formant synthesisers have been shown to produce very high quality synthesised speech. Copy synthesis, where synthesiser parameters are derived from natural speech, yields speech almost indistinguishable from the original.

HMS[7] have devised a system embodying information necessary to drive a parallel formant synthesiser from a sub-phonemic transcription. A ranked table of acoustic segments with associated formant parameters together with a set of rules defining inter and intra segment behaviour are used to convert a sub-phonemic representation to a sequence of synthesiser control frames.

The original system has a pre-set duration for each of the acoustic segments, in our system durational information is derived by the recogniser from the original speech as part of the standard recognition algorithm.

The parallel formant synthesiser can be configured with different talker tables to allow for different synthesiser voices. Although not currently implemented inclusion of speaker characteristics as part of a transmission preamble could facilitate speaker identification.

PERFORMANCE

Recognition Results

Baseline results for monophone and triphone models using 20 mel-frequency cosine coefficients (mfcc) with time differences and energy are shown in table 1. The cosine transform was on 4 kHz bandlimited 256 point FFT Hanning windowed (68 % overlap) speech data. The recogniser was trained on 39 ARM reports and tested on 10 ARM reports all from the same speaker.

	Word Error (%)	Phone Error (%)
Monophone	20.7	12.3
Triphone	10.2	5.4

Table 1: Baseline phone results

Table 2 details results for mono-segment and tri-segment models using conventional three state allocations. The same representation as above was used.

	Word Error (%)	Segment Error (%)
Mono-segment	22.4	16.0
Tri-segment	7.2	4.1

Table 2: Segment results with 3 states/model

Table 3 shows the results obtained with a modified state allocation strategy, using one state for the plosive segments.

	Word Error (%)	Segment Error (%)
Mono-segment	21.3	14.4
Tri-segment	7.4	3.8

Table 3: Segment results with modified state allocation

The above results show segment based recognition performance to be comparable to phone based recognition.

Synthesised Speech

In the absence of formal evaluation procedure, informal listening tests on quality and intelligibility were performed by the authors.

Speech quality was found to be more natural than achievable from a purely textual representation. However, it was much poorer than that achievable with a 2400 bits/sec LPC vocoder. Speaker identification is not possible on the current system since talker tables are not implemented.

Transmission Data Rates

For the ten ARM test files the average bit rate was 262 bits/sec.

CONCLUSION

A segment vocoder based on recognition and synthesis techniques has been shown to produce intelligible speech at sub 300 bits/sec. Segment based recognition performance has been shown to be comparable to conventional phoneme based recognition. An average data transmission rate of 262 bits/sec has been demonstrated.

Future work will aim to address a number of limitations of the current system.

For a task independent system the speech recogniser must behave as a phonetic transcriber explaining the incoming data purely in terms of acoustic segments without being word constrained. The system also needs to be speaker independent or speaker adaptive to cater for previously unseen speakers.

Mapping of segments to formant parameters is currently defined by a table and a set of rules. Further investigation is needed into the behaviour and derivation of the table and associated rules, in particular the possibility of automatically deriving both from natural speech.

Talker tables allow the synthesiser to be configured for different voices. Automatic derivation of speaker characteristics from a sample of speech would facilitate speaker identification.

Investigation into optimal coding of the data stream using standard data compression techniques should yield a further reduction in data rate. The effect of transmission channel errors on the synthesised speech and appropriate error protection strategies need to investigated.

REFERENCES

[1] J Picone, G R Doddington, “*A Phonetic Vocoder*”, Proc. IEEE ICASSP 89 Vol 1, pp. 580-583, 1989.

[2] V W Zue, “*Very Low Data Rate Communication*”, Technical Report, RADC-TR-84-200, Rome Air Development Center, October 1984.

[3] F K Soong, “*A Phonetically Labeled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis*”, Proc. IEEE ICASSP 89 Vol 1, pp. 584-587, 1989.

[4] K K Paliwal, P V S Rao, “*Synthesis-based recognition of continuous speech*”, J. Acoust. Soc. Am. 71(4), pp. 1016-1024, 1982.

[5] R E Donovan, P C Woodland, “*Improvements in an HMM-Based Speech Synthesiser*”, EUROSPEECH 95 Proc. Vol 1, pp. 573-576, 1995.

[6] T Masuko, K Tokuda, T Kobayashi, S Imai, “*Voice Characteristics Conversion for HMM-Based Speech Synthesis*”, ICASSP 97 Proc. Vol III, pp. 1611-1614, 1997.

[7] J N Holmes, I G Mattingly, J N Shearme, “*Speech Synthesis By Rule*”, Language and Speech, Vol 7, pp. 127-143, 1964.

[8] S R Browning, J McQuillan, M J Russell, M J Tomlinson, “*Texts of Material Recorded in the SI89 Speech Corpus*”, RSRE Research Note No. 142, February 1991.

[9] A M Noll, “*Cepstrum Pitch Determination*”,
J. Acoust. Soc. Am. 36, pp 296-309, 1964.