TOWARDS A UNIFIED MODEL FOR LOW BIT-RATE SPEECH CODING USING A RECOGNITION-SYNTHESIS APPROACH

Wendy J. Holmes

Speech Research Unit, DERA Malvern, St. Andrews Road, Malvern, Worcs.,WR14 3PS, U.K.
Tel: +44 1684 894104, E-mail: holmes@signal.dera.gov.uk

ABSTRACT

This paper proposes a recognition-synthesis approach to speech coding which uses an underlying formant trajectory model for both recognition and synthesis. It is argued that this "unified" approach to coding has the potential to achieve low data rates whilst preserving speech quality and important paralinguistic information. A simple coding scheme is described which establishes the principles of this approach. In this scheme, the formant analysis method described in [1] is first applied to the input speech. The formant features are then input to a linear-trajectory segmental hidden Markov model recognizer [2] in order to locate segment boundaries. The formant parameters for each segment are coded using a linear trajectory description, and used to drive the JSRU parallel-formant synthesizer [3] to reproduce the utterance at the receiver. The coding method has been tested on utterances from a variety of speakers. In the current system, which has not yet been optimised for coding efficiency, speech is typically coded at 600-1000 bits/s with good intelligibility, whilst preserving speaker characteristics.

1. INTRODUCTION

Successful coding of speech at low data rates of a few hundred bits/s requires a very compact, low-dimensional representation of the speech signal, which is generally applied to variable-length "segments" of speech. Automatic speech recognition is potentially a powerful way of identifying useful segments for coding. In particular, if the segments are meaningful in phonetic terms, knowledge about segment identity can be used to assist in the coding. In the extreme, very low data rates can be achieved by transmitting only phoneme identity information.

A number of recognition-based coders have been proposed which use hidden Markov models (HMMs), such as the systems described in [4-6]. In all of these systems, the coding is based on the recognition units. One possibility for reconstructing the utterance is to use the HMMs themselves. However, even with quite sophisticated schemes such as the one described in [6], the HMM assumptions of piecewise-stationarity and of independence are such that HMMs are inherently limited as speech production models. Another problem is that typical feature sets such as LPC coefficients [4] or mel-frequency cepstral coefficients [6] impose limits on the quality of the coded speech. As an alternative to HMM-based synthesis, the system described in [5] used a separate synthesis-by-rule system to regenerate the utterance, but this approach relies on the assumption that the segments identified by the HMM recognizer will also be suitable for synthesis. In all these systems, it is difficult to retain information about speaker characteristics, at least if the recognizer operates in a speaker-independent mode.

2. A "UNIFIED" SPEECH CODING MODEL

A good model for recognition-based coding needs to provide a compact representation of speech, while offering both accurate recognition performance and high quality synthesis. Such a model could be used for coding at a range of data rates, by trading bits against retention of speaker characteristics. With no limitations on vocabulary size, at the lowest bit rates speech could be generated from a phoneme sequence. At higher data rates, the coding could be applied directly to speech production parameters. This approach requires an accurate model for speech dynamics and a suitable representation of production mechanisms. Progress has recently been made towards modelling dynamics by the development of linear-trajectory segmental HMMs [2], and a useful functional representation of speech production is provided by formants. This paper focuses on a speech coding scheme that demonstrates the principle of recognition-synthesis coding using the same linear formant-trajectory model for both recognition and synthesis. The coding is applied to analysed formant trajectories, and so is at the high bit-rate end of the range of coding schemes discussed above.

3. FORMANTS FOR RECOGNITION AND SYNTHESIS

Formants have been demonstrated to provide the basis for very high quality copy synthesis of speech [7]. However, for coding applications the formant controls must be derived automatically. Unfortunately it is difficult to label formants reliably without reference to the phonetic content of an utterance, but recently a new method of formant analysis has overcome some of these difficulties when used with HMM-based recognition [1]. This formant analyser includes an estimate of confidence in measurement accuracy and, where necessary, offers alternative sets of formant trajectories for resolution in the recognition process.

In the current study the new formant analyser is used to obtain formant features for both synthesis and linear-trajectory segmental HMM recognition [2], although the confidence measure and formant alternatives have not yet been included. More work is needed to improve the formant segmental HMM recognition, but as this study represents the first attempt at also incorporating synthesis, it has concentrated on coding and synthesis aspects.

Although formants have previously been used as the basis for speech coding schemes [e.g. 8,9,10], the use of trajectory-based recognition is an important distinguishing feature of the approach proposed here. A particularly beneficial aspect is the future potential for using information from the recognition to assist in the coding by, for example, determining which of alternative possible formant trajectories is the most likely.

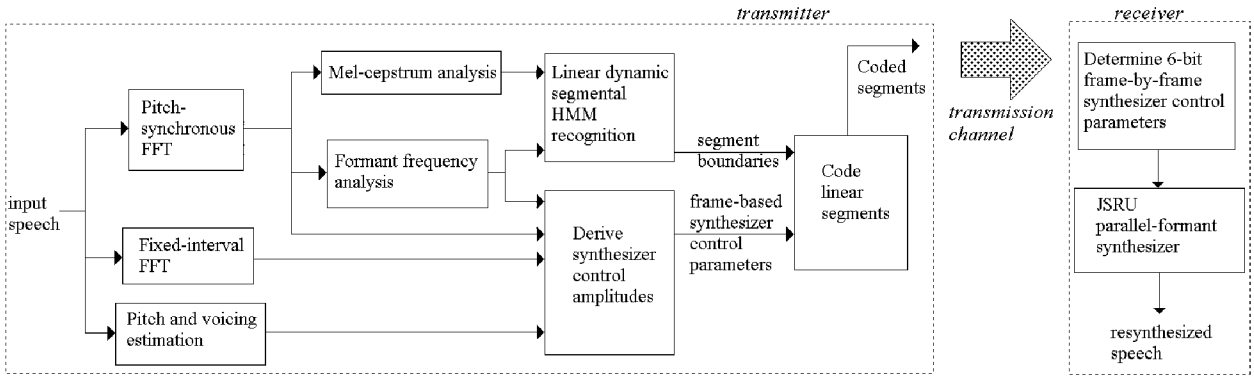

Figure 1: Block diagram of the recognition-based linear segment coding scheme.

4. A SIMPLE CODING SCHEME USING LINEAR FORMANT TRAJECTORIES

4.1. Overview of coding scheme

The major components of the coding scheme are shown in Figure 1. At the transmitter, input speech is analysed to determine formant trajectories, which are then input to a linear dynamic segmental HMM recognizer to identify segments suitable for linear-trajectory coding. By using linear formant trajectory models for recognition, the segments that are identified should be well represented by straight-line formant parameters for coding. The analysed formants are also used as the basis for deriving control parameters for formant synthesis. The synthesis is performed using the JSRU parallel-formant synthesizer [3], which has been shown to provide natural-sounding synthetic speech [7,11]. The control parameters are coded as linear trajectories for each of the identified segments. The coded segments are converted back to frame-by-frame control parameters to drive the synthesizer at the receiver.

4.2. Control parameters for JSRU synthesizer

The parameters required to control the JSRU parallel-formant synthesizer are formant frequencies and amplitudes, together with information about the excitation source. The details of the synthesizer are described in [3]. Briefly, there is a voiced and an unvoiced excitation generator, both of which are arranged to produce a spectral envelope which is substantially flat over the frequency range of the formants. The output of these generators provides the input to a parallel network of resonators with time-varying frequency and amplitude controls. The combined response of the resonators acts as a filter which shapes the excitation spectrum to model both the response of the vocal tract and the natural variation of the excitation spectral envelope. In order to generate natural-sounding synthetic speech that retains individual talker characteristics, it is therefore important to code both the formant frequency and amplitude controls fairly accurately. The 10 synthesizer control parameters for every frame (usually 10 ms) are as follows:

- Degree of voicing (V)
- Fundamental frequency (F0)
- Frequencies of the first three formants (F1, F2, F3)
- Amplitudes of these formants (A1, A2, A3)
- Amplitude of the fixed high frequency formant (AHF)
- Amplitude in the low frequency region (ALF).

4.3. Deriving synthesizer control signals

An excitation analysis program [12] is used to obtain values for fundamental frequency and degree of voicing at 10 ms intervals. The frequencies of the first three formants are provided by the output of the formant analyser described in [1]. When the analyser offers two alternative formant choices for a frame, the first choice is used. The formant amplitude controls are obtained using the FFT-based method described in [11].

The output of this processing stage is a value for each of the 10 synthesizer control parameters, specified at 10 ms intervals. Using the synthesizer in its default configuration, with six bits assigned to each of the controls, gives a data rate of 6000 bits/s. These control parameters can be used to perform frame-by-frame analysis-synthesis, and provide a useful point of reference for comparing the segmental coding results.

4.4. Segment-based coding

Recognition. Recognition is performed using phone-level linear-trajectory segmental HMMs [2]. The feature set for the recognition comprises the first three formant frequencies together with general spectral-shape information in the form of five mel-cepstrum coefficients and an overall energy feature. This feature set has previously been used successfully for conventional HMM recognition [1]. In the segmental-HMM experiments so far, the formant confidence estimates and multiple formant choices have not yet been incorporated, and hence there is considerable scope for improving the recognition performance. However, with the coding scheme presented here, provided that the recognition identifies suitable segments for linear coding, the recognition error rate does not limit the coding quality.

Each phone is modelled with an appropriate number of linear segments in order to describe its spectral characteristics, with the number of segments assigned based on phonetic knowledge. Three segments are used to model voiceless stops and affricates, with two segments being used to represent voiced stops and diphthongs. However, only one segment was considered necessary for nasals, fricatives, semivowels and monophthongal vowels. For each segment, a minimum and maximum segment duration is set to allow a plausible range of durations for each phone and keep the computation for segmental-HMM recognition at a manageable level.

Segment parameterisation. In the linear-trajectory segmental HMM, a straight line is described by its mid-point and slope. However, for the purposes of coding with a limited number of bits, a more accurate representation is provided by the segment start value and its end value expressed as the difference from the start value. This approach allows very rapid changes to be encoded, while also accurately representing gradual changes of only a few levels over several frames. The total range of trajectory slopes is such that gradual changes (which correspond to a slope value of less than unity) would have been lost if the coding had been based on a quantized slope representation.

An important advantage of transmitting segment end values expressed as differences is that the segment start value need only be transmitted for those cases where there is a sudden change (more than some specified threshold) from the value of a control signal at the end of the previous segment. This test is applied to all of the formant control signals (but not to the fundamental frequency and voicing controls). Provided that changes in all of these controls are below appropriate thresholds, only the segment end values are specified. As the synthesizer control signals change smoothly across many segment boundaries, this technique allows for a considerable saving in bit rate. In fact, it was found to be advantageous (in terms both of bit rate and of speech quality) to enforce continuity across segment boundaries involving two vowels or a vowel and a sonorant.

Initially there were some problems with the linear segment representation of the voicing and fundamental frequency, which arose when the phone segment boundaries identified in recognition did not coincide with major changes in the nature of the excitation (as excitation information did not contribute directly to the recognition process). For example, sometimes the voicing did not start until the second or third frame of a vowel segment, and there were instances in which the pitch dropped suddenly towards the end of a vowel segment due to a "creaky" voice quality. Problems caused by any excitation characteristics being incompatible with the phone segmentation were largely overcome by introducing some simple algorithms which checked for any sudden changes in the excitation characteristics. The linear model was then only fitted to regions of smooth change (with extrapolation to model the complete segment). With this approach, a linear trajectory was successfully used to code the voicing and fundamental frequency for most speech segments, in addition to being appropriate for the formant controls.

Bit allocation To code a segment, it is necessary to include its duration, together with straight-line parameters for each of the synthesizer controls. The longest allowed segment duration is set at 16 frames, with the result that the segment duration can be coded using four bits. The voicing control (V) need only be represented very coarsely, which is achieved with the four levels provided by a two-bit control. Five bits are used for each of F0, F1, F2, ALF, A1 and A2, with four bits each for F3 and A3 and only three bits for AHF. This bit allocation is similar to the one described in [8] for a variable frame-rate coding scheme.

For segments longer than a single frame, an additional flag bit is needed to indicate whether or not start values are specified as well as end values. The total numbers of bits are as follows:

Smooth-join segments	
Duration	4
Flag bit	1
F0,V x 2	14
F1,F2,ALF,A1,A2	25
F3,A3	8
AHF	3
Total	55

Segments with an abrupt change	
Duration	4
Flag bit	1
F0,V x 2	14
F1,F2,ALF,A1,A2 x 2	50
F3,A3x2	16
AHF x 2	6
Total	91

For any single-frame segments, 47 bits are required.

It should be noted that with this coding scheme the bit rate does not depend on the recognition vocabulary, but it does depend on the number of segments identified per second of speech. Factors affecting the number of segments include speaking rate and acoustic complexity of the words in the vocabulary.

5. EXPERIMENTS

5.1. Experimental method

The coding method has been tested on the speaker-independent connected-digit recognition task used in earlier speech recognition experiments [1,2], and also on a speaker-dependent task of recognizing spoken airborne reconnaissance mission (ARM) reports using a 500-word vocabulary. For each task, the formant analyser [1] was applied to the training data. Sets of linear-dynamic segmental HMMs were then trained using the method described in [2], but with a feature set comprising the three formant frequencies together with five cepstral coefficients and an overall energy feature. The coding scheme described in Section 4 was then applied to a variety of utterances from the test sets for each of the recognition tasks. For each utterance coded, the bit rate was calculated (excluding any regions of silence) and the quality of the coding was evaluated by informal listening tests. The coded speech was compared with the original natural utterance and with the frame-by-frame analysis-synthesis on which the coding was based.

5.2. Coding Results

For most of the utterances tested, a good approximation to the original six-bit frame-by-frame synthesizer control signals was provided by coding using the linear segments identified from recognition with the bit allocation described in Section 4.4. Although the controls were somewhat quantized, particularly for the higher formants, all the main characteristics of the original control signals were preserved in the segment coding. Even the very simple formant-based linear segmental HMM used here was found to be generally effective at identifying linear trajectories for coding, regardless of recognition errors.

Listening to the speech, the frame-by-frame analysis-synthesis (at 6000 bits/s) generally produced a very close copy of the original natural speech. The synthetic utterance was distinguishable from the original with careful listening but, when played in isolation, sounded acceptable as a recording of natural speech. Formant analysis errors caused occasional problems, which in most cases would have been avoided by using the second choice provided by the formant analyser. Extending the segmental-HMM recognizer to identify the correct formant choice would provide the necessary information.

The segment-coded utterances generally sounded more stylised than the frame-by-frame analysis-synthesis. However, the main characteristics of the original speech were preserved, as can be seen from the spectrograms shown in Figure 2. The segment coding scheme generally produced speech that was highly intelligible and retained speaker characteristics for all of the speakers tested. In some cases the synthetic speech actually benefited from the smoother quality provided by the segmental coding scheme in comparison with the frame-based synthesis.

For the digit data, typical coding rates were 600-800 bits/s. For the ARM task, which included more acoustically-complex words and for which the reports were spoken quickly, the rates tended to be higher at about 800-1000 bits/s. These rates reflect the nature of the speech material, and **not** the vocabulary size.

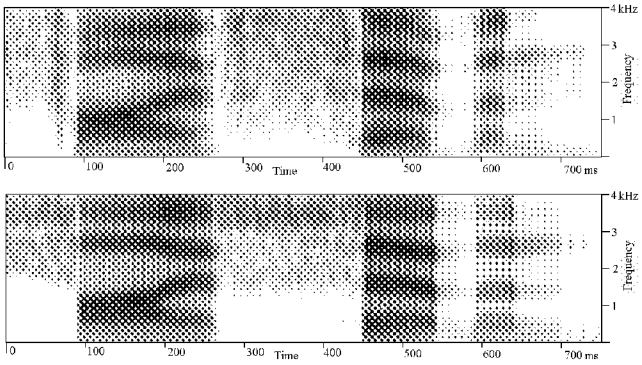

Figure 2: Spectrograms of original natural speech (top) and segment-coded version (below) of an utterance of "five seven".

6. CONCLUSIONS

The experiments described here have demonstrated the potential of a recognition-synthesis coding scheme using a linear-trajectory formant model for both recognition and synthesis. A simple approach, in which the recognizer is used to identify segments which are then coded as linear synthesizer control parameters, has been shown to achieve good quality coded speech at rates of less than 1000 bits/s, with speaker characteristics clearly preserved.

Within the current coding scheme, there are a number of areas for further work. Possibilities include the following:

1. Some saving in bit rate should be possible by careful reduction of the bits assigned to many of the synthesizer control parameters. Alternatively, larger reductions may be possible by applying vector quantization techniques.
2. The recognition performance of the formant segmental HMMs should be improved by including the confidence measure and evaluation of alternative possible formant trajectories within the recognition. These developments could use similar techniques to those that have been found to be successful for conventional-HMM recognition [1].
3. The quality of the coding would be improved by reducing any formant analysis problems. Information from the recognition can be used to assist in the analysis of formant trajectories for the coding. Once the recognition has been extended to evaluate alternative formant trajectories, the outcome of this process could be used to dictate which of the two possible sets of formant trajectories to transmit for coding. It may also be beneficial to use the confidence measure to guide the coding by, for example, not coding variations in low-confidence formant frequencies.

The quality of the speech coding should improve as the recognition and synthesis stages become more integrated within a common framework. Future developments could achieve lower bit rates by progressing towards a truly unified model which would allow good quality synthesis from the recognition models themselves.

7. ACKNOWLEDGEMENTS

Thanks are due to Martin Russell for helpful early discussions about this project, and to John Holmes for making available his excitation analysis program.

8. REFERENCES

1. Holmes, J.N., Holmes, W.J. and Garner, P.N. "Using formant frequencies in speech recognition", *Proc. EUROSPEECH'97*, Rhodes, pp. 2083-2086, 1997.
2. Holmes, W.J. and Russell, M.J. "Linear dynamic segmental HMMs: variability representation and training procedure, *Proc. IEEE ICASSP'97*, Munich, pp. 1399-1402, 1997.
3. Holmes, J.N. "A parallel-formant synthesizer for machine voice output", in *Computer Speech Processing*, F. Fallside and W.A. Woods (Eds.), Prentice-Hall International, 1985.
4. Picone, J. and Doddington, G.R. "A phonetic vocoder", *Proc. IEEE ICASSP'89*, Glasgow, pp. 580-583, 1989.
5. Ismail, M. and Ponting, K. "Between recognition and synthesis – 300 bits/second speech coding", *Proc. EUROSPEECH'97*, Rhodes, pp. 441-444, 1997.
6. Tokuda, K., Masuko, T., Hiroi, J., Kobayashi, T. and Kitamura, T. "A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques", *Proc. IEEE ICASSP'98*, Seattle, pp. 609-612, 1998.
7. Holmes, J.N. "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", *IEEE Trans. Audio and Electroacoustics*, 21, pp. 298-305, 1973.
8. McLarnon, E., Holmes, J.N. and Judd, M.W. "Experiments with a variable-frame-rate coding scheme applied to formant synthesizer control signals", *Proc. Speech Communication Seminar*, Stockholm, pp. 71-79, 1974.
9. Dupree, B.C. "Formant coding of speech using dynamic programming", *Electronics Letters*, 20, pp. 279-280, 1980.
10. Zolfaghari, P. and Robinson, T. "A segmental formant vocoder based on linearly varying mixture of Gaussians", *Proc. EUROSPEECH'97*, Rhodes, pp. 425-428, 1997.
11. Holmes, W.J. "Copy synthesis of female speech using the JSRU parallel formant synthesizer", *Proc. EUROSPEECH'89*, Paris, pp. 513-516, 1989.
12. Holmes, J.N. "Robust measurement of fundamental frequency and degree of voicing", *Proc. ICSLP'98*, Sydney, 1998.