Proceedings of the Institute of Acoustics

LOW BIT-RATE SPEECH CODING USING A LINEAR-TRAJECTORY FORMANT REPRESENTATION FOR BOTH RECOGNITION AND SYNTHESIS

Wendy J. Holmes
Speech Research Unit, DERA Malvern,
St Andrews Road, Malvern, Worcs WR14 3PS, UK

ABSTRACT

This paper presents a recognition-synthesis approach to speech coding which uses an underlying formant trajectory model for both recognition and synthesis. The formant analysis method described in [1] is first applied to the input speech and the formant features are then input to a linear-trajectory segmental hidden Markov model (HMM) recognizer [2]. The segment boundaries located by the recognizer are used as a basis for linear-trajectory segment coding of the analysed formant trajectories. The paper describes a number of special features of the approach which are aimed at making the most effective use of the formant analyser output, and focuses on the use of information from the recognition process to guide the coding. The resulting coded formant trajectories are 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 speech coding at low data rates of a few hundred bits/s requires a 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. If the segments are meaningful in phonetic terms, knowledge of segment identity can be used to guide the coding. In the extreme, very low data rates can be achieved by transmitting only the phoneme identities.

A number of recognition-based coders have been suggested that use HMMs, such as the systems described in [4-6]. In these systems, 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 they 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 coded speech quality. 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 are suitable for synthesis. In all these systems, it is difficult to retain information about speaker characteristics, at least if the recognizer is used in a speaker-independent mode.

In [7], it is proposed that the above issues can be addressed by a "unified" approach to speech coding in which the same (appropriate) model of speech production is used as the basis for both the recognition and synthesis. The principles of this approach are demonstrated by a simple coding scheme based on linear formant trajectories. The recognition uses linear-trajectory segmental HMMs [2] applied to formant features produced by the formant analyser described in [1], and the synthesis uses the JSRU parallel-formant synthesiser [3]. The current paper describes further developments which are aimed at making the most effective use of the formant analyser output for both recognition and synthesis, and in particular focuses on the use of information from the recognition process to assist in the formant trajectory coding. Although formants have been used as the basis for other speech coding schemes [e.g. 8, 9, 10, 11], the use of trajectory-based recognition is an important distinguishing feature of the approach being pursued here.

2. RECOGNITION USING LINEAR FORMANT TRAJECTORIES

2.1. The formant analyser

Although formant frequencies are known to be important in determining the phonetic content of speech sounds, they are not generally used as features for automatic speech recognition. Formant features are difficult to extract reliably reliably without reference to the phonetic content of an utterance, and one or more of the formants may be poorly defined in some sounds (such as fricatives). Recently a new method of formant analysis [1] has overcome some of these difficulties by including provision for dealing with ambiguous labelling and with indistinct formants. To deal with indistinct formants, where any one estimate of formant frequency is likely to be unreliable, each formant frequency estimate is assigned a value representing confidence in its measurement accuracy. In cases of ambiguity, the formant analyser offers two alternative sets of formant trajectories for resolution in the recognition process.

An example of the analyser output is shown in Figure 2, to demonstrate the use of alternative sets of formant trajectories. The analyser provides labels for the first three formants, but this speaker has considerable energy in F4 (often more than in F3). As a result, two alternative sets of trajectories have been offered in all three vowels, where the incorrect choice has used F3 to model what is in fact F4, and omitted one of the other formants. In the first vowel, the second choice is a better approximation, but in the other two vowels the first choice is correct.

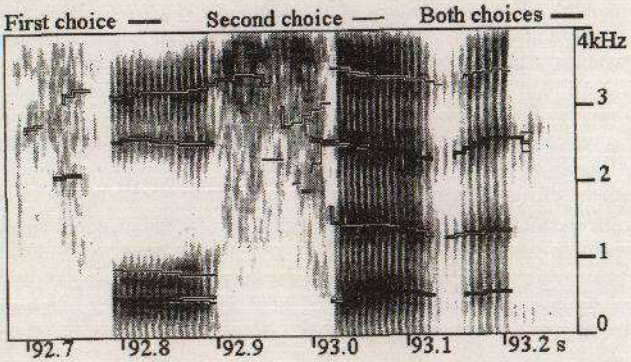

Figure 1: Spectrogram of an utterance of the words "four seven", with superimposed formant tracks showing alternative formant allocations offered by the analyser for F1, F2 and F3. Tracks are not plotted when there is no confidence in their accuracy.

2.2 Using the formant analyser output for recognition

The output of the formant analyser has previously been sucessfully used for conventional HMM recognition [1,12], with a feature set comprising the first three formant frequencies together with general spectral-shape information in the form of five mel-cepstrum coefficients and an overall energy feature. Alternative formant sets were accommodated by simply choosing the formant set which gave the highest HMM emission probability for each frame and model state. In [12], it is shown that, by representing the formant confidence values as variances (large variance represents low confidence), the confidence measures can be rigorously incorporated into HMM probability calculations. For recognition, the result is simply that the formant confidence variance is added to the model variance. Thus low-confidence features have very little influence on the recognition decision.

2.3. Linear-trajectory segmental HMM recognition using formant features

Linear-trajectory segmental HMMs [2] address limitations of conventional HMMs as models of dynamics by associating each model state with a linear trajectory to describe underlying trends in how features change over time. Trajectory models are particularly well suited to modelling formant trajectories which

are inherently smoothly changing. The formant confidence variances can be used in the segmental-HMM probability calculations in an extension of the approach previously used for conventional HMMs. In the case of segmental HMMs, the effect is that the high-confidence formants within any one segment are the ones that have most influence on the segment probabilities. The alternative formant trajectories have been used by finding the set of trajectories that gives the highest probability over the entire segment when computing any one segment probability. The ability of the segmental HMMs to naturally enforce continuity of a formant choice within a segment is an important advantage over a simple frame-based decision which allows conventional HMMs to swap between the two alternatives on successive frames. The segmental approach used here may occasionally be too restrictive, as it does not allow for any cases where valid trajectories may be obtained by using one choice in one part of a segment and the other choice in another part of the segment. However, the approach is generally effective as changes in trajectory choices usually occur at regions of large spectral change, which tend to coincide with segment boundaries.

3. CODING LINEAR FORMANT TRAJECTORIES

Formants have been demonstrated to provide the basis for very high quality copy synthesis of speech [13]. However, for coding applications the formant controls must be derived automatically. This process suffers from the same tendency to formant extraction problems as was discussed above in relation to recognition. However, by using the new formant analysis method described in the previous section in conjunction with the recognition stage, the coding approach described here addresses the formant extraction problems.

3.1. Overview of coding scheme

The major components of the coding scheme are shown in Figure 2. At the transmitter, input speech is analysed to determine formant trajectories, which are then input to a linear dynamic segmental HMM recognizer. The recognizer acts to identify segments suitable for linear-trajectory coding. As part of the recognition process, a choice is made between any alternative sets of formant trajectories, and the selected formants are used as the basis for deriving control parameters for formant synthesis. Linear trajectory parameters are estimated, taking into account the confidence estimates and the segment identities as determined by the recognizer (see below for more detail). Synthesis is performed using the JSRU parallel-formant synthesizer [3], which has been shown to provide natural-sounding synthetic speech [13,14]. 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.

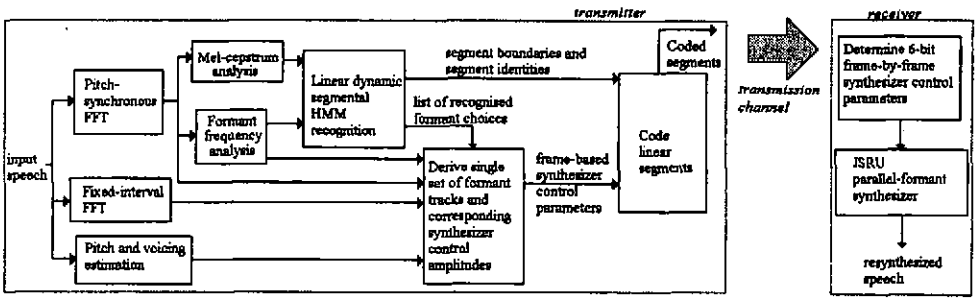

Figure 2: Block diagram of the new recognition-based linear segment coding scheme.

3.2. 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 vocal tract response 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).

3.3. Deriving synthesizer control signals

An excitation analysis program [15] 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 there are alternative trajectories, those chosen by the recognizer are used. The formant amplitude controls are obtained using the FFT-based method described in [14]. 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 segmental coding results.

3.4. Segment-based coding

Recognition. Recognition is performed using phone-level linear-trajectory segmental HMMs as described in Section 2. 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, affricates and some diphthongs, with two segments being used to represent voiced stops, most diphthongs and a few long monophthongs. However, only one segment was considered necessary for nasals, fricatives, semivowels and most 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.

Fitting linear trajectories. For each synthesizer control signal in each segment, the best straight line fit to the frame-by-frame values is determined using a least mean square error criterion. For the formant frequencies, the error for each frame is weighted by the appropriate confidence variance, so that the most reliable frames have the largest influence. For segments during which none of the estimates for a formant are reliable, it may be better not to use the formant values produced by the analyser at all. Possibilities being investigated are to simply interpolate between the trajectories for adjoining segments, or in some cases to use pre-specified phone-dependent values. These techniques help to avoid nasty effects due to spurious sudden formant movement at segment boundaries, as well as helping to reduce the bit rate.

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.

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 if 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 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, speech quality was found to be improved by **forcing** continuity of formant frequencies across segment boundaries involving two vowels or a vowel and a sonorant.

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 that 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		Segments with an abrupt change	
Duration	4	Duration	4
Flag bit	1	Flag bit	1
F0,V x 2	14	F0,V x 2	14
F1,F2,ALF,A1,A2	25	F1,F2,ALF,A1,A2 x 2	50
F3,A3	8	F3,A3x2	16
AHF	3	AHF x 2	6
Total	55	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.

4. EXPERIMENTS

4.1. 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 the feature set and special treatment of formant features as described in Section 2 above. The coding scheme described in Section 3 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 a simple frame-by-frame analysis-synthesis.

4.2. 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 3.4. 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. The formant-based linear segmental HMM was found to be generally effective at identifying linear trajectories for coding.

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, usually sounded acceptable as a recording of natural speech. When synthesizing using the first-choice formant trajectories, formant analysis errors caused occasional problems, as can be seen from the example shown in Figure 3. In many cases, such as in the example of Figure 3, these problems were then avoided in the segmental coding, as the recognizer correctly used the second choice provided by the formant analyser.

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. 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. There were also cases (such as the one shown in Figure 3) for which the ability to use the recognizer to assist in the selection of alternative formant trajectories allowed the segment coding to be better than a simple frame-by-frame analysis-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.

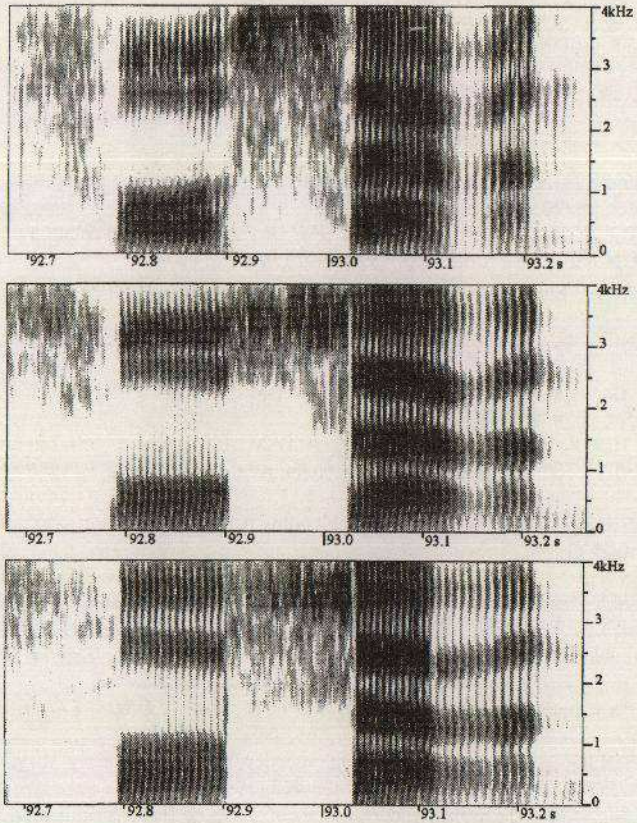

Figure 3: Coding the utterance "four seven" based on the formant analysis shown in Figure 1. The natural speech (top) is shown coded on a frame-by-frame basis (middle) using the first-choice formant trajectories, which failed to detect F2 during the vowel in "four". Using the segment-based coding (bottom), the recognizer selected the second-choice formant trajectories for this region and these are used for the synthesis, so providing a better approximation to the correct F2 position. During the other vowels, the recognizer correctly used the first choice formant trajectories and these are therefore used for both the frame-by-frame and the segment-coded versions.

5. 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. Information from the recognition has successfully been used to assist in the coding process and so to improve over the system described in [7]. Good quality coded speech has been achieved at rates of less than 1000 bits/s, with speaker characteristics clearly preserved. 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.

In the long term, it should be possible to achieve better quality speech coding, together with better recognition performance and more accurate formant analysis, by further integrating all these aspects 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.

6. 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.

7. REFERENCES

[1] J N HOLMES, W J HOLMES and P N GARNER, 'Using formant frequencies in speech recognition', *Proc. EUROSPEECH'97*, Rhodes, pp. 2083-2086 (1997)

[2] W J HOLMES and M J RUSSELL, 'Linear dynamic segmental HMMs: variability representation and training procedure', *Proc. IEEE ICASSP'97*, Munich, pp. 1399-1402 (1997)

[3] J N HOLMES, 'A parallel-formant synthesizer for machine voice output', in *Computer Speech Processing*, F. Fallside and W.A. Woods (Eds.), Prentice-Hall International (1985)

[4] J PICONE and G R DODDINGTON, 'A phonetic vocoder', *Proc. IEEE ICASSP'89*, Glasgow, pp. 580-583 (1989)

[5] M ISMAIL and K PONTING, 'Between recognition and synthesis – 300 bits/second speech coding', *Proc. EUROSPEECH'97*, Rhodes, pp. 441-444 (1997)

[6] K TOKUDA, T MASUKO, J HIROI, T KOBAYASHI and T KITAMURA, 'A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques', *Proc. IEEE ICASSP'98*, Seattle, pp. 609-612 (1998)

[7] W J HOLMES, 'Towards a unified model for low bit-rate speech coding using a recognition-synthesis approach', *Proc. ICSLP'98*, Sydney (1998)

[8] E McLARNON, J N HOLMES and M W JUDD, 'Experiments with a variable-frame-rate coding scheme applied to formant synthesizer control signals', *Proc. Speech Communication Seminar*, Stockholm, pp. 71-79 (1974)

[9] B C DUPREE, 'Formant coding of speech using dynamic programming', *Electronics Letters*, 20, pp. 279-280 (1980)

[10] N SEDGWICK, 'Emulation of a formant vocoder at 600 and 800 BPS', *Proc. EUROSPEECH'93*, Berlin, pp. 523-526 (1993)

[11] P ZOLFAGHARI and T ROBINSON, 'A segmental formant vocoder based on linearly varying mixture of Gaussians', *Proc. EUROSPEECH'97*, Rhodes, pp. 425-428 (1997)

[12] P N GARNER and W J HOLMES, 'On the robust incorporation of formant features into hidden Markov models for automatic speech recognition', *Proc. IEEE ICASSP'98*, Seattle, pp. 1-4 (1998).

[13] J N HOLMES, '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)

[14] W J HOLMES, 'Copy synthesis of female speech using the JSRU parallel formant synthesizer', *Proc. EUROSPEECH'89*, Paris, pp. 513-516 (1989)

[15] J N HOLMES, 'Robust measurement of fundamental frequency and degree of voicing', *Proc. ICSLP'98*, Sydney (1998)

© British Crown Copyright 1998 / DERA

Published with the permission of the controller of Her Britannic Majesty's Stationery Office