

SEGMENTATION OF THE SPEECH WAVEFORM USING DIFFERENTIATED FIRST  
AND SECOND FORMANT TRACKS

A.P Breen.\*

ABSTRACT

This paper will introduce a technique for segmenting speech about points of zero gradient using differentiated first and second formant tracks. Three LPC and formant tracking algorithms were examined, and the most reliable combination used in the next stage of the analysis. The resulting formant tracks were then modelled by quadratics.

Finally, the results produced by applying the analysis procedure to two sets of speech data are given and some conclusions drawn.

INTRODUCTION

ASR systems which attempt to recognise a word or sentence on the basis of their phonetic content, segment the signal as a first stage in this process. Some systems attempt to locate, and phonetically label the segments in one step, while others, locate segments which are then labelled at the next stage.

The criteria adopted to segment speech vary. Some methods locate segments at constant intervals of time e.g 10ms, from which parameters are extracted. These small segments are bundled together to create larger segments which are then phonetically labelled (ref 1). Other methods extract parameters such as voicing and aspiration from the waveform. Empirical rules are then applied to these parameters to form segments (ref 2).

This paper will introduce a method of acoustic segmentation based on the assumption that, points of zero gradient in differentiated formant tracks reflect the achievement or partial achievement of acoustic targets in their production of a phonetic element.

It is proposed that segments located about these points represent a natural way to segment speech, that they will reflect the phonetic content of the speech, and that parameters extracted about these points are relatively resistant to speaker and occasion variability.

OUTLINE OF STUDY AND METHOD

The technique of segmentation using differentiated formant tracks requires good formant data; care was therefore taken in deciding which method of LPC and which formant tracker should be used in this procedure. Initially, three methods of LPC were considered: fixed frame, larynx synchronous, and closed phase larynx synchronous LPC (ref 3). Spectrograms were produced from each of the three methods. These were then compared with each other and with a filter bank spectrogram of the same utterance, to determine which of the LPC analysis methods produced the most reliable formant estimates.

Larynx synchronous LPC seemed to produce the most reliable formant estimates for the data considered, i.e. isolated words spoken by a female speaker.

----

\*Dept. of Phonetics and Linguistics, University College, London.

Fixed frame analysis was inferior in quality in all cases considered, while closed phase larynx synchronous LPC occasionally produced erroneous formant estimates.

The formant estimates produced were then formant tracked. Initially, three formant trackers of differing complexity were considered, the best of which, produced by GEC Hirst for the SPAR project was used in the next stage of the analysis. This formant tracker divides the utterance up into voiced frames which are then fed to a labelling function. This likelihood function is maximised via dynamic programming (ref 4).

Fig.1 shows the differentiated first formant of the word "zero", spoken by a female RP speaker, and calculated by simply differencing the formant track. The track produced is clearly too noisy for this application. To overcome this problem the formants produced by the formant tracker were simplified using quadratic modelling. The modelled formants were constructed by piece-wise fitting the quadratics over a given number of frames. To reduce the possibility of transitional effects at quadratic segment boundaries, the differentials are at present calculated by finding the gradient of the best straight line fitted about the point and two points either side.

Fig.2 shows three instances of formant tracks for the sentence "The pearl was worn in a thin silver ring", recorded under anechoic conditions by a male RP speaker. The top display was produced by the formant tracker, while the middle display was produced by modelling the formants every 10 frames and the bottom display by using quadratics every 30 frames. Fig.2 demonstrates that an increase in the number of frames over which the quadratic is fitted simplifies the formant tracks.

Fig.3 shows the differentiated formant tracks for the same sentence using formants modelled every 30 frames. Once the differentiated formant tracks have been produced the points of zero gradient are found. Segmentation points are found for both the first and second differentiated formant tracks. The segmentation points produced by the differentiated first formant are then compared with those produced by the differentiated second formant and any segmentation points which lie within 20ms of each other are combined into one point. The remaining segmentation points represent the segmentation for that utterance.

The speech data considered in this paper consists of (a) three repetitions of the word "zero" recorded under anechoic conditions spoken by a female RP speaker and (b) the sentence "The pearl was worn in a thin silver ring" spoken under anechoic conditions by a male RP speaker and in a relatively noisy environment by a male non RP speaker. Each of the data items were transcribed and the duration of phonetic segments determined by a trained phonetician. For example the word "zero" was transcribed:

/z/	120ms - 280ms
/rɪə/	280ms - 440ms
/r/	440ms - 490ms
/əʊ/	490ms - 700ms

Each segmentation point created by the segmentation procedure was labelled with the phonetic transcription associated with that time. At present the procedure creates segmentation points either side of a voiceless region. When this occurs it is the practice to label the segmentation points bordering the voiceless region as belonging to the phonetic segments abutting that area and an extra segmentation point is created at the centre of the voiceless region.

## RESULTS

All the results given in this section exploited segments produced by modelling the formant tracks with quadratics every 30 frames. The phonetic transcription produced by the segmentation procedure, for the first occurrence of the word "zero" was:

/zzɪəɪəɪəʊəʊ/

while the transcription for the second occurrence was:

/zzɪəɪəɪəʊəʊ/

and the transcription for the third occurrence was:

/zzzɪəɪəɪəʊəʊ/

The original transcription for the sentence "The pearl was worn in a thin silver ring" was:

/ðə pɜ:l wəz wɔ:n ɪn ə θɪn sɪlvə rɪŋ/

However, the transcriptions produced from the segmentation procedure for the RP and non RP speakers were:

/ððə pɜ:ɜ:ɜ:l wəɛz wɔ: n ə θɪnnnn sɪlv rɪŋ/

and

/ððə pɜ:ɜ: wəz wɔ: n ə θɪnnn sɪlv rɪŋŋ/

## DISCUSSION

The results from the experiments suggest that while it is acceptable to model isolated words with quadratics every thirty frames, segment deletion occurs when this amount of modelling is applied to sentences.

Quadratic modelling enables any required amount of formant detail to be deleted by simply increasing the number of frames over which the quadratic is constructed. However, it is clear from the above results that a point will be reached where the model is no longer adequately reflecting the characteristics of the formant track. When this point is reached, segments will be missed or spurious segments created. If the formant tracks are modelled by quadratics over too few frames the technique will produce too many segments. A compromise will inevitably have to be reached between over production of segments and segment deletion.

## REFERENCES

1. D.R. Reddy et al, "A Model and a System for Machine Recognition of Speech", IEEE Trans. Audio Electroacoust., vol. AU-21, pp. 229-238, June 1973.
2. C.J. Weinstien et al, "A System for Acoustic-Phonetic Analysis of Continuous Speech", IEEE Trans. Acoust., Speech, and Signal process., vol. ASSP-23, pp. 54-67, Feb. 1975.
3. D.J.B Pearce and L.C Whitaker, "Reference Formant Analysis", IEE Int. Conf. Proc. on Speech Input/Output; Techniques and Applications. pp. 37-43, 1986.
4. K. Frimpong-Anash. Pending Publication.

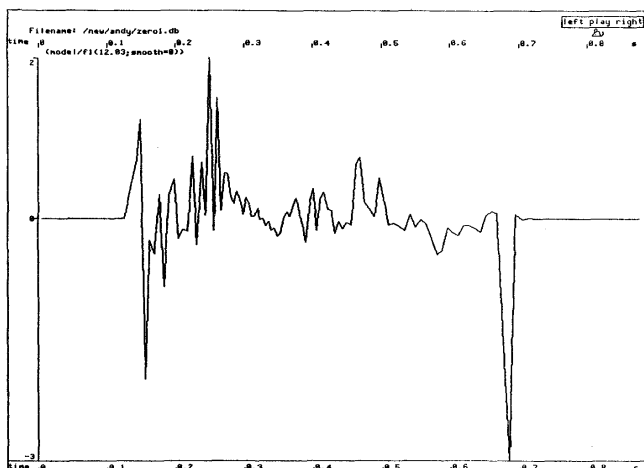


Fig.1  
Differentiated  
first formant for  
the word "zero".

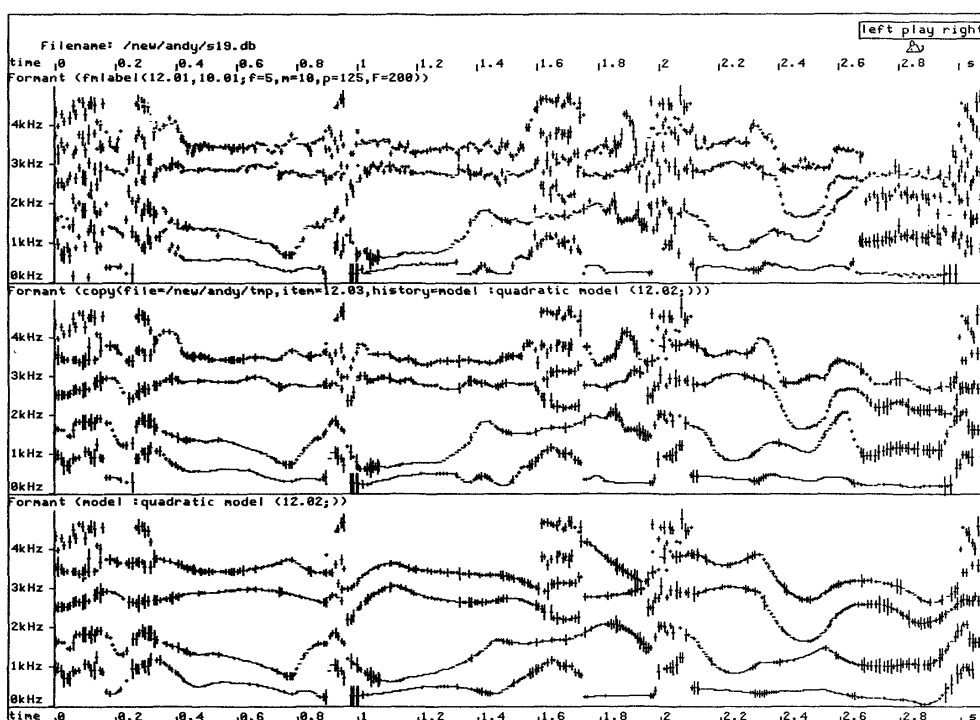


Fig.2 First four formant tracks. Top, formant tracker. Middle, 10 frame quadratic modelling. Bottom, 30 frame quadratic modelling.

Fig.3  
Differentiated first  
and second formants using  
quadratic modelling.

