Proceedings of The Institute of Acoustics

AN ENDPOINT DETECTION ALGORITHM INCORPORATING ACOUSTIC-PHONETIC KNOWLEDGE

I.G. Fletcher, E. Rooney, F. McInnes, M.A. Jack

Centre for Speech Technology Research, University of Edinburgh

INTRODUCTION

Accurate separation of background noise from speech segments in streams of isolated words is important to many areas of speech research.

(a) Pattern-matching techniques for comparing input words with reference templates are greatly affected by accuracy of boundary location. (Wilpon et al [1] have shown how crucial it is in isolated word recognition.)

(b) The computation involved in analysis of the isolated word is greatly reduced if all the silent periods are removed.

(c) In large-scale statistical analysis of phonetic features, storage can be reduced, and automatic identification/extraction/analysis of phonetic features may be possible.

In many applications the identity of the word is known in advance but the endpoint detection algorithm used is one intended for automatic word recognition where clearly the input word is unknown. This paper discusses how knowledge of the phonetic characteristics of the particular word may be introduced into an established endpoint detector to improve the reliability and accuracy of endpoint detection in non-ideal conditions for those applications with a priori knowledge of word identity. The next section describes the speech database used in this work and problems commonly encountered in endpoint detection. Following this is a description of the basic endpoint detection algorithm used. Methods of improving accuracy and reliability of endpoint detection using word classification are then outlined. Finally, a summary of continuing work and of the limitations caused by inter speaker variability is given.

DATABASE USED AND PROBLEMS ENCOUNTERED IN ENDPOINT DETECTION

The database used consisted of a representative subset of a larger set of 390 British speakers. The vocabulary consisted of randomly ordered sequences of all the digits, ('0' - '9'), and the letters 'g' ,'m' ,'n' ,'s' ,'t', 'w', 'x', 'y', and 'z', repeated three times during each session. The recording sites were relatively quiet rooms within office accommodation. Sources of background noise were mainly non-stationary: typically lorries and aircraft from outside, background speech and ringing telephones, inside; and air-conditioning, The signal to noise ratio was very variable due to the nature of the noise. No automatic gain control or noise cancellation systems were employed. DC offset was calculated using 0.1 secs. of silence. The speech was low-pass filtered to 8 kHz, sampled at 20kHz, with a 12-bit analogue to digital conversion. The measures used: signal magnitude (*sig. mag.*) and zero-crossing rate (*zcr*), were calculated every 10 msecs., with a 10 msec. frame.

The problems encountered were those of general low amplitude coupled with large amplitude variations, caused by nervousness. These effects are less prevalent in later recordings, where speakers are used to the equipment and have been advised

about optimal signal level.

THE BASIC ENDPOINT DETECTION ALGORITHM

The original endpoint detection algorithm was based on work by Rabiner et al [2], and Lamel et al [3], but with empirically based adaptive noise thresholds. The state diagram of Figure 1. serves to illustrate the operation. The signal magnitude sm(t) at time t. is defined as:

$$sm(t) = \sum_{n=N(t-1)}^{Nt} S(n) \quad ; N = framelength. \tag{1}$$

NT1,NT2 are the adaptive noise thresholds, initially set to NT1 = 7000, and NT2 = 18000 (average signal magnitude = 35 and 90). They are then adapted using the empirically based adaptation equation:

$$NT1 = NT1 + A\_RATE((sm(t) - NT1/18000)(NT1 + 3000)) \tag{2}$$

$$NT2 = NT2 + A\_RATE((sm(t) - NT1/45000)(NT1 + 2000)) \tag{3}$$

with adaptation rate, A_RATE = 0.1. The adaptation ensures that the thresholds change with varying noise conditions, which were prevalent in the recording environment.

A pulse is not accepted unless its level is above a threshold level PT for a time greater than MIN_HI = 4 consecutive frames. PT = 25000 (average signal magnitude 125). If a pulse falls below NT1 for MAX_LO consecutive frames, it is rejected. MAX_LO is set equal to 2 frames. FT_MAX and RT_MAX are the rise time and fall times of the pulses: both set at 4 frames. The minimum gap between accepted pulses, MIN_GAP, is set at 280msecs..

After pulses have been identified, they are checked for length: minimum of 70msecs., maximum of 850msecs.. The start and end of these, possibly concatenated, pulses are the provisional word boundaries. It should be noted that there is no pulse ordering [1], as this would require storage of the whole speech sequence.

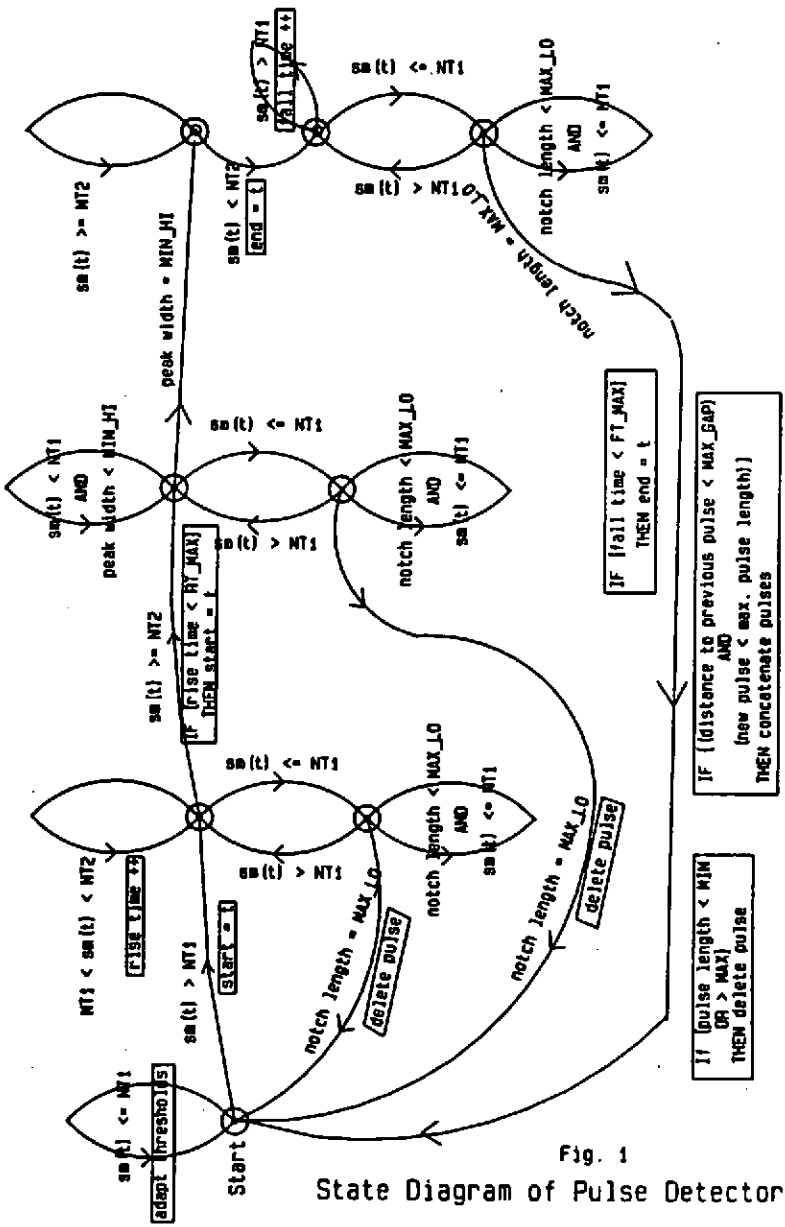
A search is then carried out, up to 250 msecs.. beyond the provisional endpoint, for high *zcr* regions. If, in this period, there are ZMIN (equal to 4) frames with a *zcr* greater than ZT1 = 20 crossings (equivalent to 1kHz), then the word boundary is extended to the frame, with *zcr* above this threshold, which is furthest from the provisional endpoint.

HEURISTIC IMPROVEMENT USING WORD CLASSIFICATION

Original endpoint detector weaknesses

Weak initial fricatives such as the /f/ in 'four' and 'five', were sometimes missed because the region of high *zcr* was too short, when measured from the boundary of the energy pulse. These segments were also missed due to the weakness of the *zcr* increase.

AN ENDPOINT DETECTION ALGORITHM



Notes: 1. At each transition, t is incremented.
2. All actions are boxed

Fig. 1
State Diagram of Pulse Detector

In words such as 'six' and 'x', where there was a marked distance in time between the consonant and vowel, and the energy in the consonant was too low to be picked up by the *sig. mag.* detector, errors were possible.

The release burst of stop consonants were occasionally missed due to being too short, even though the region of high *zcr* was quite pronounced.

The 'skirt' of a *sig. mag.* pulse was often inaccurate because of the conservative thresholds.

Lip smacks before and after words were often included within the word boundary.

Algorithm development

The task here is to find word classifications which lead to algorithms specific to particular sets of phonetic characteristics, and hence improve the quality of endpoint detection for the majority of speakers uttering the words which include that characteristic. In parallel, we have to ensure that the quality of endpoint detection for particular aberrant speakers is not degraded.

Our approach was to identify all the words which had failed on the initial endpoint detection. If the failure was attributable to some phonetic characteristic of that word not being detected, attempts were made to classify the characteristic using the *zcr* and *sig. mag.* measures. An algorithm which catered for this particular class was then devised. The criterion for deciding on the success of a new algorithm was, simply, did it improve the endpoint detection, assessed manually, without degrading it in any case. The other failure was where a non-speech segment was wrongly within the word boundary. An improvement was expected by ensuring that all phonetic characteristics at word boundaries had some particular suitable algorithm. Thus sounds which fell outside the expected classifications would not be detected.

The crude (because of wide inter-speaker variation) classes which have been used to correct the weaknesses highlighted above, are given below with corresponding solutions.

Class I. If the word is known to have an endpoint with no associated high frequency component (As in words which start or finish with pure voiced sounds), then a new endpoint based on more detailed use of the original *sig.*mag. can be found. This can be done by taking a 3-point average of the energy measure and extending the word boundary, while the filtered sm(t) remains above NT1, AND it's gradient does not change sign.

The remaining classes use the *zcr* measure. A search space from 3 frames inside the provisional endpoint to a period ZEXT frames beyond it is first defined. A threshold, ZT1 is then drawn across the *zcr* curve within this space. The points at which pulses of *zcr* cross threshold ZT1 are now found. If there are no pulses the endpoint extension is abandoned. Two pulses are combined if less than MAX_ZGAP frames apart.

Class II. If a pronounced broad region of high *zcr.* is expected the search space is 90 frames, and MAXZ_GAP equals 3 frames. A check is made that there are more

than (1.5 * Z_MIN) points above ZT1 (in this case equal to 20 frames). If there are not, the extension is abandoned. If there are, the biggest is found by summing each of the *zcr* values within the pulse. The endpoint is extended to the last point in this pulse.

Class III. If a pronounced narrow region of high *zcr* is expected the search space is 90 frames, ZT1 equals 20, and MAXZ_GAP equals 3 frames. No check is made for the number of points above ZT1 being > Z_MIN. However, a check is made that the biggest pulse has at least one point greater than 1.5 * Z_T1. If so, the endpoint is extended to the past point in this pulse.

Class IV. If a weak fricative is expected the search space is 45 frames, MAXZ_GAP equals 6 frames, and ZT1 equals 18. The lower threshold and more relaxed concatenation requirement allow less pronounced areas of high *zcr* to be included. If there are more than (1.5 * Z_MIN) points above ZT1, the endpoint is extended to the biggest pulse. If this condition is not met but the *zcr* pulse beginning is very close (less Z_MIN + 1 frames) to the provisional endpoint, AND there is at least one point greater than (ZT1 * 1.5), then the endpoint is extended to this pulse.

Class V. This is a non-specific 'catch-all' class. The same course is followed as for Class IV, except the search space is 30 frames, MAXZ_GAP equals 3 frames, and ZT1 equals 20. The *zcr* pulse is only "close" to the provisional endpoint if less than (Z_MIN/2 + 1) frames away.

After this final extension a further check on word length may be carried out.

Using these class-specific algorithms, there was a marked improvement in perceptually evaluated endpoint detection. There was a similar improvement in the performance of a speaker verification system which used isolated words. Figures 2. and 3. show *sig.* *mag.* and *zcr* for the words six and eight, with endpoints marked.

CONCLUSIONS

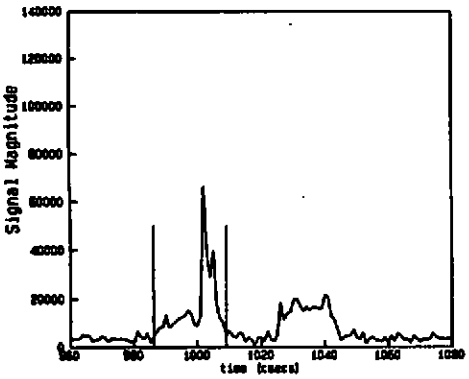
This initial study has shown that, for certain applications, the use of stored acoustic-phonetic knowledge in endpoint detection can be valuable. The technique may be improved by a better choice of parameters and a more comprehensive range of algorithms. Lamel [3] and Wilpon [1] use a log energy array normalised to the background noise level, which seems to provide good performance and may remove the need for the, computationally intensive, adaptive noise thresholds. The following parameters, some of which have been tried by Wilpon [1], are currently being investigated for suitability in producing word classifications/algorithms.

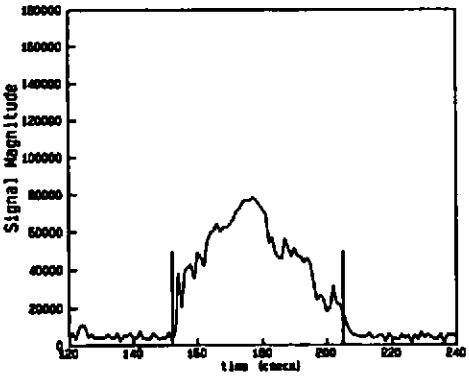
Log energy array.

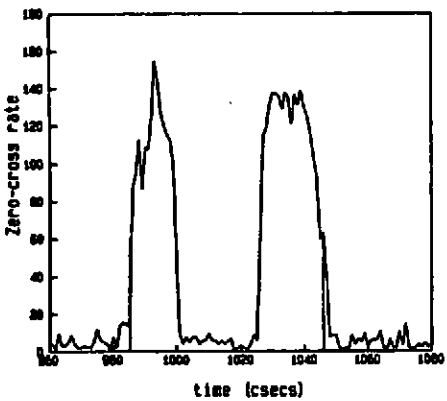
- Pulse width and height.
- Gap between pulses.
- Slope of pulse at start and end.
- Relative size of adjacent pulses.

Zero-crossing rate.

AN ENDPOINT DETECTION ALGORITHM







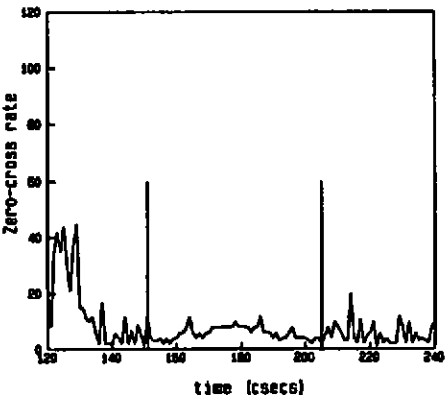


Figure 2.
Plots of sig. mag. and zero-cross rate for the word six.

Figure 3.
Plots of sig. mag. and zero-cross rate for the word nine.

Pulse width and height.
Gap from energy pulse.
Concatenation rules for pulses.
Relative size of adjacent pulses.

It may be possible to reject word-like noise such as coughs, by ensuring that the expected phonetically characteristic features come within a certain range, and in a particular order.

Any classifications have to take into account the inter speaker variation in producing the phonetic features in question. A rigorous investigation of such variations and their effects on the proposed algorithms will need to be carried out. The effects of various British accents on the word classification have already been noted. For instance, the final release in /n`in/, and the trilled /r/ in four.

REFERENCES

1. J. Wilpon, L.R. Rabiner, and G. Martin, 'An improved word-detection algorithm for telephone quality speech incorporating both syntactic and semantic constraints', AT&T Bell Labs. Tech. J., Vol. 63, (1984).
2. L.R.Rabiner and M.R. Sambur, 'An algorithm for determining the endpoints of isolated utterances', B.S.T.J., Vol. 54, no.6, 297-315, (1975).
3. L.F. Lamel, L.R.Rabiner, and A.E. Rosenberg, 'An improved endpoint detector for isolated word recognition', I.E.E.E. Trans. A.S.S.P., Vol. ASSP-29, 777-785, (1981).