THE ACOUSTIC BASIS OF THE DISTINCTION BETWEEN STRONG AND WEAK VOWELS

Jonathan Harrington

Speech, Hearing, and Language Research Centre, Macquarie University, Sydney, Australia

ABSTRACT - This study is a preliminary exploration of the acoustic-phonetic basis of the distinction between 'strong' and 'weak' vowels. Segments were taken from a database of continuous speech and were classified using critical band and formant frequency parameters. The results show that up to 75% of strong vowels and just over 83% of weak vowels are correctly classified, depending on the type of acoustic classification used.

INTRODUCTION

It will come as no surprise to Psycholinguists studying human word recognition that a major difficulty for any machine recogniser of continuous speech is the automatic detection of word boundaries and the segmentation of the acoustic speech signal into words. Although the acoustic cues for the separation of minimal phrases such as 'grey tie'/'great eye' are well documented (Lehiste, 1960), there is still insufficient acoustic-phonetic knowledge to enable most word boundaries to be automatically detected on a purely bottom up basis. Instead, word boundaries are hypothesised by matching a phonemic or phonetic input directly to the lexicon (Harrington & Johnstone, 1987), or else the phonetic/phonological segments which are recognised are subjected to pre-lexical processing. In Church (1983) for example, some word-boundaries are derived by allophonic parsing while in Lamel & Zue (1984) they are obtained from a knowledge of three-phoneme sequence constraints which occur across word boundaries but not word-internally (see also Harrington, Watson & Cooper, 1989).

The problem with a recognition model in which word boundary identification depends on narrow, or very narrow (Church, 1983) phonetic units is that such units are themselves very difficult to recognise. An alternative strategy may be to resolve the speech waveform into much coarser units, such as broad (Huttenlocher & Zue, 1984) or mid (Dalby, Laver & Hiller, 1986) classes, which can be much more reliably detected, and then to derive word boundaries from this coarser representation, without analysing the speech signal into narrower phonetic units. However, this approach may be unworkable because broad or mid classes do not contain a sufficient amount of information for word boundaries to be located in continuous speech (Harrington & Johnstone, 1987; Harrington *et al.*, 1989).

A possible way around this impasse is suggested by a human model of word recognition, developed recently by Cutler in a number of papers (Cutler & Norris, 1988) which provides the mechanism for the detection of many word boundaries from a fairly coarse segmentation of the speech signal into 'strong' and 'weak' vowels. Central to this model is the perception of word boundaries at strong syllable onsets in English. In support of this model, Cutler & Norris (1988) have demonstrated that listeners are slower to detect the embedded word 'mint' in /minteiv/ than in /mintəʃ/. This slower reaction time is predicted by the strong syllable onset model because /teiv/ is strong and /təʃ/ is weak: accordingly, a word boundary is perceived before the former, thus destroying the phonetic integrity of 'mint', but not before the latter. A further argument in support of the strong syllable onset model is based on the phonological structure of words in English. Studies by Cutler & Carter (1987) and Harrington *et al.* (1989) have shown over 70% of words in English have initial strong syllables, this percentage being a good deal higher when word frequency is taken into account (Cutler & Carter, 1987). These statistics suggest that listeners would actually perceive a very large proportion of word boundaries correctly if their perceptions were guided by strong syllable onsets. Compatibly, Harrington *et al.* (1989) have shown that over 40% of word boundaries are correctly located in 145 transcriptions of continuous speech by a strategy which is similar in most respects to Cutler's strong syllable onset model. Importantly from the point of view of automatic speech recognition, a hit-rate of over 40% is also obtained when the transcriptions are encoded as broad classes. Thus the relevance of the strong syllable onset model to the machine recognition of continuous speech is that a large number of word boundaries should be detectable from a fairly coarse phonetic representation of the speech signal.

There are however a number of implicit acoustic-phonetic assumptions in Cutler's word recognition model which need to be addressed in extending the model to include word boundary detection by machine. Firstly, there is the problem of defining 'strong' and 'weak' vowels. In Cutler & Norris (1988), the distinction seems to be motivated largely be lexical/phonological considerations (syllables with primary and secondary stress are always strong) but also sometimes by the phonetic properties of the vowel which is strong if it is 'full' (i.e. peripheral in the phonetic vowel space) but weak if it is central. Thus, the vowel of the first syllable in 'fantastic', which is lexically unstressed in some frameworks (e.g. Fudge, 1985) can be strong if it is realised as [a] but weak if it is realised as a more central vowel. Quite how the exact cutoff is to be specified on what is possibly a continuum from [a] to [ə] that is affected by paralinguistic factors of tempo and 'formality' of speech style, remains unresolved. Perhaps more fundamentally, it is not even clear whether the separation of strong and weak vowels is feasible in many cases. For instance, the durational and spectral differences between the vowels of 'hid' (strong) and 'ability' (both weak), or between 'foot' (strong) and 'annual' (weak except perhaps in extreme citation form) are likely to be minimal in most cases, or at least fraught with the same difficulties which confront the automatic recognition of very narrow phonetic units.

The experiment reported in this paper is a preliminary investigation into the second of these issues, that is the extent to which strong and weak vowels can be separated at an acoustic-phonetic level of analysis. The procedure in this experiment is to characterise strong and weak vowels in a multidimensional acoustic space based on an analysis of part of a digitised and hand-labelled database of spoken British English (RP). The results that are reported are for the classification of strong and weak vowels from a different part of the same speech database, based on their proximity to these multidimensional spaces. Results are also reported for the classes *approximant consonant* and *nasal consonant* because these are classes with which strong and weak vowels are also likely to be confused: clearly, the strong syllable onset model of word recognition is only of value in the machine recognition of speech, not only if strong and weak vowels can be separated from each other, but also if these two categories are distinguishable from other major segment classes.

METHOD

Speech segments

The segments were taken from a database collected at the Centre for Speech Technology Research, Edinburgh University (CSTR) between 1984 and 1989. At the time of the analysis, the database included 98 phonetically dense sentences (e.g. "I'm naming one man among many") each produced twice by four male speakers of British English, Received Pronunciation (RP). The utterances were recorded in a sound-treated recording studio and digitised with 12-bit resolution at 16 kHz on a Masscomp computer. All utterances were subsequently segmented and labelled by trained Phoneticians using digital spectrographic and transcription facilities of the speech signal processing package *Audlab*. The segmentation divided the acoustic speech signal approximately into phoneme size units (Williams & Dalby, 1987) with corresponding phonemic labels assigned. A subsequent conversion of the labelling was necessary to distinguish strong from weak nuclei. Strong nuclei included all nuclei with primary and secondary stress (e.g. 'manner', 'perpendicular, 'analogue') of all content words and of one or two function words which are always produced with a full vowel (e.g. 'while'). The remaining nuclei were labelled as weak. The 98 utterances produced twice by the four male speakers were divided into two groups, the *training data* and the *classification data*. The training data consisted of all strong vowels (n = 1994), all weak vowels (n = 1427), all approximant consonants (n = 1045) and all nasal consonants (n = 939) that occurred in two readings of 60 utterances each produced by the four speakers. The classification data included all strong vowels (n = 1173), all weak vowels (n = 944), all approximant consonants (n = 745) and all nasal consonants (n = 518) that occurred in two readings of the remaining 38 utterances each produced by the same four speakers.

Acoustic parameters

There were four different conditions which depended on whether training and classification were

carried out using formant data or critical band data, and whether the data were obtained from the midpoint or from three different points in the segment. In the first condition, *Fm* (Formant midpoint), the first three formant centre frequencies and their bandwidths were obtained for each segment at the segment's midpoint using an automatic formant tracking procedure developed by Crowe & Jack (1987). In the second condition, *Ft* (Formant, three points), formant frequencies and their bandwidths were obtained at each segment's 20%, 50% (midpoint), and 80% time points. In the third condition, *Cm* (Critical band, midpoint), energy values in the first 18 critical bands (corresponding to a frequency range of 0.2 - 6.4 kHz) were obtained at each segment's midpoint. Following Chang & Cheung (1986), the first critical band had a lower cutoff at 200 Hz in an attempt to exclude fundamental frequency information. Amplitude variation was normalised by subtracting the broadband energy in the 0.2 - 6.4 kHz range from the energy values in each of the critical bands (Chan & Cheung, 1986; Klein *et al.*, 1970). In the fourth condition, *Ct* (Critical bands, three points), amplitude normalised critical bands in the same frequencies were obtained at the 20%, 50% and 80% time points of each segment. The total segment duration was also included as a parameter in each of the four conditions.

In summary, there are 7 dimensions in the *Fm* condition (3 formant centre frequencies, 3 formant bandwidths, segment duration), 19 dimensions in the *Ft* condition (3 formant centre frequencies x 3 points, 3 formant bandwidths x 3 points, segment duration), 19 dimensions in the *Cm* condition (18 critical bands, segment duration) and 55 dimensions in the *Ct* condition (18 critical bands x 3 points, segment duration).

Training and classification

For each of the four conditions separately, training consists of three stages: speaker-normalisation, dimension reduction using discriminant analysis, and the calculation of class centroids and covariance matrices.

Speaker-normalisation consists of subtracting a speaker's centroid from the training data. A centroid in this case is a vector of mean values, one for each dimension. For example, in the *Fm* condition, a speaker centroid is a 7 dimensional vector of the means of the first three formant centre frequencies, first three bandwidths and segment duration calculated over the segments of the training data produced by the same speaker. Subtracting speaker centroids in this way has been shown to be an effective and simple way of removing speaker-specific effects (Chan & Cheung, 1986; Klein *et al.*, 1970).

Discriminant analysis is included in the training stage as a data reduction technique for reducing the high dimensional space (e.g. 55 dimensions in condition *Ct*). Essentially, discriminant analysis eliminates the redundant information that arises due to dimensions being correlated with each other (Klecka, 1980). In the first stage of discriminant analysis, a set of coefficients, which are weightings on the original dimensions, are derived and are used to transform the data to n - 1 dimensions, where n is the number of different classes. Since in this case there are four classes (strong vowels, weak vowels, approximant consonant, nasal consonant), the high dimensional spaces in each of the four conditions can be transformed maximally to three new dimensions.

When a new segment is classified, it must first be speaker-normalised by subtracting the speaker's centroid derived from the prior training stage of analysis. Additionally, it must also be transformed to a three-dimensional space using the coefficients which were derived at the training stage using discriminant analysis. Once the speaker-normalisation and data reduction procedures have been carried out, the Mahalanobis distance is calculated in the three-dimensional space from the new segment to each of the class centroids obtained in the training stage of the analysis. The new segment is classified as one of the four classes strong vowel, weak vowel, approximant consonant or nasal consonant, based on whichever Mahalanobis distance is the shortest. All the statistical calculations in this paper were achieved using the software package 'Acoustic Phonetics in S' (Watson, 1989) which enables calculations to be made on the acoustic parameters relative to the start and stop times of the segments assigned in hand-labelling the data.

RESULTS

The results of the classifications in the four different conditions are shown in Tables 1 - 4. Turning firstly to the formant data, the results show that 73% of strong vowels, 44.2% of weak vowels, 80.5% of approximant consonants and 73.9% of nasal consonants are correctly classified in the *Fm* condition, in which formant frequencies and bandwidths were obtained only at the midpoint. The results of the *Ft* condition, in which formant centre frequencies and bandwidths are additionally obtained at the 20% and 80% points, show an improvement for all categories except approximant consonants. In particular, there is a much higher correct classification score for weak vowels (71%) in the *Ft* condition compared with the *Fm* condition (44.2%).

Turning now to the critical band data, the results show correct classification scores of 71.9% for strong vowels, 75.1% for weak vowels, 71.8% for approximant consonants and 75.5% for nasal consonants in the *Cm* condition, in which energy values in critical bands were obtained at the midpoint. The corresponding scores in the *Ct* condition, in which critical band data was also obtained at three time points, are 75% correctly classified for strong vowels, 83.3% for weak vowels, 75.8% for approximant consonants and 81.7% for nasal consonants: these correct classification scores are higher in all four classes compared with the scores in the *Cm* condition.

A comparison of the formant data (Tables 1 and 2) with the critical band data (Tables 3 and 4) shows that better classification scores are generally obtained from critical bands. When formant data and critical band data are obtained only at the midpoint (Tables 1 and 3 respectively), the classification scores for the classes strong vowel, approximant consonant and nasal consonant are quite similar, but there is a much higher correct classification score (75.1%) for weak vowels in the *Cm* condition compared with the *Fm* condition (44.2%). When formant data and critical band data are obtained at three points (Tables 2 and 4), higher correct classification scores are obtained in the *Ct* condition for three categories (weak vowel, approximant consonant, nasal consonant) while classification scores for strong vowels are very similar in both conditions.

DISCUSSION

The main aim of this study has been to carry out a preliminary investigation of the acoustic basis for the separation between strong and weak vowels. In the best of the four conditions examined, 75% of strong vowels and 83.3% of weak vowels were correctly classified while 22.8% of strong vowels were misclassified as weak and 10.5% of weak vowels were misclassified as strong. In the same condition, less than 0.5% of strong or weak vowels were misclassified as nasals and less than 6% of either strong or weak vowels were misclassified as approximant. The study is an open test in that different segments were used in the training and classification stages and it is also semi-speaker independent because the multidimensional distributions for the four classes are based on speaker-normalised segments from all four speakers together, rather than from any single speaker.

The results of the study suggest that more reliable scores are obtained from critical band, compared with formant frequency and formant bandwidth data. The reasons for this are still unclear. It may be that the automatic formant tracking procedure produced unreliable values in a small number of cases, or else that critical bands contain better discriminatory information for the separation of the four major classes in this study. The study also shows that better classification scores are obtained when parameter values are obtained at three time points rather than just the midpoint. It may be possible to relate this finding to a number of perceptual and acoustic studies by Strange and colleagues (Strange, 1989) which show that the transitions into and out of vowels are as important as the acoustic vowel target for their identification. At this stage, such an interpretation must be made with caution in view of the fact that the data in this study were obtained for the midpoint, which does not always coincide with the vowel target.

An explanation for the misclassifications between the four major classes must await the outcome of a more detailed evaluation. However, a further analysis of some of the data has shown that one of the main sources of confusion is due to the substantial acoustic variability of schwa. Additionally, many of the confusions betweens strong and weak vowels are caused by the variable classification of /i/ and /i:/: thus strong vowels in 'hit', 'hid' etc. were often classified as 'weak' while weak vowels

in e.g.'city' were often classified as 'strong'. This latter misclassification may be due to an increasing tendency in RP to produce final /i/ as a long, rather than a short, vowel, which would cause it to be phonetically more similar to strong vowels.

In summary, the study has provided some experimental evidence that a large proportion of strong and weak vowels are acoustically distinguishable. The study is currently being extended to include more speakers and a larger database. It may also be necessary to review the definitions of 'strong' and 'weak' in the light of a more detailed analysis of the misclassifications that were produced in this study.

REFERENCES

Chan L. & Cheung Y. (1986) Analysis and recognition of isolated Putonghua vowels by Karhunen-Loeve transformation techniques. *Speech Communication*, 5, 299-330

Church K. (1983) *Phrase-Structure Parsing: A method of Taking Advantage of Allophonic Constraints. Indiana University Linguistics Club*: Indiana

Crowe A. & Jack M. (1987) Globally optimising formant tracker using generalised centroids. *Electronic Letters*, 23, 1019-1020

Cutler A. & Norris D. (1988) The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Performance and Perception*, 14, 113-121

Cutler A. & Carter D. (1987) The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Dalby J., Laver J. & Hiller S. (1986) Mid-class phonetic analysis for a continuous speech recognition system. In Lawrence R. (ed.) *Proc. of the Institute of Acoustics*, 8, 461-469. Institute of Acoustics: Edinburgh

Fudge E. (1985) *English Word-Stress*. George Allen & Unwin: London

Harrington J. & Johnstone A. (1987) The effects of equivalence classes on parsing phonemes into words. *Computer Speech and Language*, 2, 273-288

Harrington J., Watson G. & Cooper M. (1989) Word boundary detection in broad class and phoneme strings *Computer Speech and Language*, 3, 367-382

Huttenlocher D. & Zue V. (1984) A model of lexical access based on partial phonetic information. *Proceedings ICASSP*, 26.4.1-26.4.4

Klein W., Plomp R. & Pols L. (1970) Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America*, 48, 999-1009

Klecka W. (1980) Discriminant analysis. In Sullivan J. (ed.) *Sage University Paper Series on Quantitative Applications in the Social Sciences*. Sage: Beverley Hills and London.

Lamel L. & Zue V. (1984) Properties of consonant sequences within and across word boundaries. *Proceedings ICASSP*, 42.3.1-42.3.4

Lehiste I. (1960) An acoustic-phonetic study of internal open juncture. *Phonetica*, 5, Supplement.

Strange W. (1989) Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, 85, 2135-2153.

Watson G. (1989) An environment for acoustic phonetic research (abstract). *Journal of the Acoustical Society of America*, 85, S56.

Williams B. & Dalby J. (1987) Segmentation criteria for EUSIP data base. Unpublished manuscript, CSTR, Edinburgh University.

ACKNOWLEDGEMENTS

This research was carried out while the author was a research fellow at CSTR, Edinburgh University and was part of the Alvey funded, CSTR Edinburgh/G.E.C. Marconi project in automatic speech recognition (SERC grant nos. GR/D29604, GR/D29628, GR/D29661).

TABLE 1

Classified as:	Strong	Weak	Approx.	Nasal
Input segments:				
Strong	73.0	10.5	15.9	0.6
Weak	9.9	44.2	42.3	3.7
Approx.	3.5	12.1	80.5	3.9
Nasal	2.3	2.7	21.0	73.9

Classification scores for *Fm* condition

TABLE 2

Classified as:	Strong	Weak	Approx.	Nasal
Input segments:				
Strong	75.6	16.1	7.7	0.5
Weak	12.7	71.0	15.5	0.8
Approx.	6.4	17.5	73.4	2.7
Nasal	1.7	10.0	13.9	74.3

Classification scores for *Ft* condition

TABLE 3

Classified as:	Strong	Weak	Approx.	Nasal
Input segments:				
Strong	71.9	21.0	7.0	0.1
Weak	11.0	75.1	13.8	0.1
Approx.	5.8	21.2	71.8	1.2
Nasal	5.2	5.4	13.9	75.5

Classification scores for *Cm* condition

TABLE 4

Classified as:	Strong	Weak	Approx.	Nasal
Input segments:				
Strong	75.0	22.8	2.0	0.2
Weak	10.5	83.3	5.9	0.3
Approx.	6.6	16.5	75.8	1.1
Nasal	2.9	5.2	10.2	81.7

Classification scores for *Ct* condition