EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

AUTOMATIC IDENTIFICATION OF PHONEME BOUNDARIES USING A MIXED PARAMETER MODEL

Paul Micallef, Dept. of Communications and Computer Engineering, University of Malta.
Ted Chilton, School of Electronic Engineering, Information Technology and Mathematics,University of Surrey, UK.
E-Mail: pjmica@eng.um.edu.mt; E.Chilton@ee.surrey.ac.uk

ABSTRACT

The identification of phoneme boundaries in continuous speech is an important problem in areas of speech recognition and synthesis. The use of robust parameters to allow a trained data set obtained from one language to be used for boundary identification in another language is being investigated. In particular the use of mixed time-frequency rate parameters, and the training on the change of the rate parameters at acoustic boundaries is reported.

10.21437/Eurospeech.1997-164

INTRODUCTION

The identification of phoneme boundaries in continuous speech is an important problem in areas of speech recognition and speech synthesis. In particular, speech synthesis requires accurate knowledge of phoneme transitions, in order to obtain a naturally sounding speech waveform from stored parameters. Recently [1] , [2] there has been detailed analysis reported in the literature on automatic alignment and segmentation of speech data. In [1] the features used are the cepstral parameters, while in [2] the features are auditory filter bank parameters and energy. The methods involve the use of HMM's or neural networks to identify phonemes or phoneme classes. In this paper the use of acoustic features instead of parametric features, for automatic segmentation is investigated. Al-Hashemy, [3], uses non parametric features for the discrimination of speech data. His results showed that a mixture of time and frequency features in one vector give better results than considering each feature separately. Recently, [4], [5], the use of mixed parameter sets has been reported for phoneme boundary identification. Different parameters provide differing indications of boundaries for different phoneme transition classes. The use of mixed parameters appears to give very robust boundary identification and, in particular, works well on untried speech data after being optimised with a known, labelled speech training base [6]. The work reported here, centres on the development of a speech synthesis system for the Maltese language. While the methods employed will be applicable to any language, the testing of automatic phoneme boundary detection requires an expansive, annotated speech data base which, as yet, does not exist in the Maltese language.Thus the method has been developed using the TIMIT database but with the expectation that, with the use of mixed parameters, the method is robust enough to be applied to another language, in this case Maltese, in which the phoneme boundaries are unknown.

MIXED PARAMETER MODEL

Here, a combination of both time and frequency domain parameters were used. The zero-crossing count, log energy and the first (unit sample delay) auto correlation coefficient were employed as time parameters. For the frequency domain representation, a spectral envelope was obtained on a frame-by-frame basis and the average energy output of a set of mel-scale filters gave a measure of the distribution of spectral energy. Each parameter was normalised over the utterances under analysis and then the *rate of change* of each parameter was calculated. At each boundary it is assumed that a given parameter can increase, decrease or remain the same, while between boundaries the parameters are approximately constant. The rate of change of each parameter therefore exhibits changes at a boundary. These changes are usually very robust and independent of the speaker and speaker variation. By using mixed features that are not LPC based, the rate of change at the boundary of the rate parameters has acoustic interpretation. Seven broad classes were used. These are silence, vowel, stop closure, stop burst, voiced fricatives, unvoiced fricatives and sonorants. In this way all classes of phoneme transition were adequately represented. Using the labelled training data, obtained from the TIMIT database, the phonemic symbols were transformed into the broad category sets. For each possible combination of categories the appropriate boundaries according to

burst - sonorant

	Z	C	P	L	M	H
DPK	-0.10	0.08	0.16	0.07	0.06	-0.03
KXL	-0.16	0.17	0.32	0.29	0.26	0.01
ARC	-0.14	0.11	0.30	0.41	0.21	0.01
GAR	-0.23	0.18	0.34	0.68	0.52	0.03
FILE	-0.19	0.15	0.41	0.51	0.36	0.00

vowel - voiced fricative

	Z	C	P	L	M	H
DPK	0.37	-0.50	-0.42	-0.65	-0.75	0.67
KXL	0.55	-0.40	-0.19	-0.16	-0.17	0.54
ARC	0.44	-0.66	-0.50	-0.09	-0.04	0.94
GAR	0.61	-0.42	-0.22	-0.20	-0.13	1.62
FILE	0.38	-0.44	-0.23	-0.31	-0.21	0.77

fricative - vowel

	Z	C	P	L	M	H
DPK	-0.64	0.52	0.55	0.91	0.61	-0.17
KXL	-0.56	0.63	0.35	0.62	0.35	-0.26
ARC	-0.31	0.26	0.58	1.10	1.00	0.02
GAR	-0.95	0.31	0.47	0.72	0.63	0.00
FILE	-0.52	0.56	0.47	0.73	0.54	-0.23

closure - burst

	Z	C	P	L	M	H
DPK	0.24	-0.16	0.21	0.10	0.13	0.48
KXL	0.07	-0.22	0.17	0.02	0.12	0.20
ARC	0.08	-0.26	0.25	0.13	0.15	0.41
GAR	0.01	-0.05	0.16	0.04	0.07	0.25
FILE	0.12	-0.13	0.20	0.09	0.14	0.49

Table 1

Average change of rate parameters at (a) burst - sonorant boundary; (b) vowel - voiced fricative boundary; (c) fricative - vowel boundary; (d) closure - burst boundary; The values are for individual speakers from the TIMIT training set, and for the overall FILE average for the speakers used from the training set.

the data base, were located. The values of the rate of change parameter sets were averaged for several neighbouring frames of speech data spreading across a boundary. For each boundary class, the distribution of the rate of change parameters depends on the class type. Table 1 gives values obtained for rate of change at a boundary for different speakers in the TIMIT training set, and the overall averages across a large speaker base. Using the maximum likelihood method [7], the parameter sets, derived for each located boundary from the training data, were used to build a covariance matrix for a particular phoneme transition from data representing that transition. Thus for each class of phoneme transition, one optimum covariance matrix was stored representing that class. Sentences from the training set were used to build up the matrices.

Therefore the system does not have phoneme trained or phonemic class trained data. It has transition trained data, and the covariance matrices are essentially optimised transition matrices, expressing the variability in the rate of change of the chosen parameters across the class boundary in the immediate area of the class boundary.

EXPERIMENTAL RESULTS

For the purposes of testing, another set of labelled TIMIT data was used initially. Two types of tests were conducted. The first was to check on the reliability of the matrices to obtain a correct transition frame. Parameters were derived from this test data in the same manner as for the training data for a series of 12 frames on either side of a phoneme boundary, representing an overall 120 ms of waveform. It is noted that, in this case the class of transition is known *a priori* but what is required is the position of that transition. Each frame of parameters is matched to the appropriate covariance matrix and a set of likelihood values are obtained. The minimum log likelihood c_{opt}, is obtained using

$$c_{opt} = \min[(\underline{p}-\underline{\mu}_{\underline{p}})^{T} C_{\underline{p}}^{-1} (\underline{p}-\underline{\mu}_{\underline{p}})] \quad (1)$$

where $\underline{p}$ is the vector under test, $C_{\underline{p}}$ is the covariance matrix for the boundary class, and $\underline{\mu}_{\underline{p}}$ is the mean vector for the class. The frame of test parameters which give the minimum c_{opt} is taken as the position of the boundary of the phoneme transition. Results gave a very high incidence of boundary identification within 20 ms of the TIMIT labelling. Results for some classes are shown in Table 2. The second test was to look at the discrimination between the various classes. The test was set up as a confusion matrix type in which a given known transition type parameters were submitted to all the matrices representing the various class boundaries.

The automatic segmentation was then tested on TIMIT test sentences, changing the TIMIT phonetic labels to the phonetic classes. The purpose was to see whether matching and identification of phoneme boundary classes was possible. In this case the errors could not have insertions or deletions or substitutions, but only deviations of the boundaries. For each boundary all the prospective positions,

Boundary Type	Correct	Total	%
Vowel - Fricative	789	831	96 . 1
Fricative - Sonorant	130	145	89 . 6
Silence - Vowel	96	110	87 . 2
Silence - Voiced Fricative	231	277	83 . 4
Burst - Sonorant	493	600	82 . 1
Vowel - Voiced Fricative	459	571	80 . 4
Closure - Burst	1554	2024	77 . 1
Vowel - Sonorant	1753	2440	71 . 2
Burst - Voiced Fricative	21	32	65 . 6
Sonorant- Vowel	1527	2522	60 . 5

Table 2
Boundaries within 20 ms of the annotated boundary type for various classes of boundaries.

obtained as minima from (1) were stored. A dynamic programming algorithm was then used to choose the optimal path by backtracking. Duration information on each phoneme class was included as a weighting in calculating the distance of every possible minimum from the current boundary to every possible minimum in the next boundary. This is given by

$$\min \sum_{n=1}^{N} \sum_{j=1}^{L} D_n(j) \qquad (2)$$

where N is the number of boundaries
L the number of feature frames

and

$$D_j(n) = d_n(j) + \min \sum_{i=0}^{j} d_{n-1}(i) + d_w(j,i)$$

where $d_n(j)$ is the value obtained from (1) for the frame, $d_{(n-1)}$ (i) is the minimum value at frame i from the previous boundary calculation; $d_w(j,i)$ is a distance weighting between frames j and i that depends on the expected phone type between frames i and j . To reduce the computation, only the values and position of the minima obtained from (1), at every boundary calculation, are kept as potential boundary points, the other frames being set to an arbitrary high value.

The method described here was also used with twelve lpc derived cepstral parameters rate of change as a frame vector. The tests were to see the robustness of the training with respect to the test sentences, and the performance with respect to the cepstral rate parameters. Table 3 shows the results obtained on a subset of the TIMIT test sentences. The mixed parameter set has a slightly better performance in boundary identification compared to the cepstrally derived parameters. It is also quite robust with very little difference between the result obtained using training sentences and test sentences.

TIMIT Set	Features	% Correct
Training	AP	63 . 1
	Cepstral	60 . 2
Test	AP	62 . 9
	Cepstral	57 . 6

Table 3
Results obtained for Acoustic Rate Parameters and Cepstral rate Parameters using TIMIT.

Finally, the method was applied to unlabelled Maltese utterances to obtain the phoneme boundaries and extract the diphones from the embedded words. The results appear very encouraging. Figure 1 shows a typical result of using the trained matrices to automatically segment the word 'sebveta' to get the phoneme pair /bv/ .

CONCLUSIONS

It has been shown that the use of acoustic parameters and, specifically the rates of change of these parameters, provides a robust method for both boundary identification and phoneme transition class identification where the trained data set has been obtained from a language different from that under

test. Such methods are also robust when tested on speakers unknown to the training set.

REFERENCES

[1]C G Jeong, H Jeong, "Automatic phone segmentation and labeling of continuous speech", Speech Communication, v20 pp291-311, 1997.

[2]A.Vorstermans, J P Martens, B Van Coile, "Automatic segmentation and labelling of multi-lingual speech data", Speech Communication, v 19 pp 271-293, 1997

[3]B A R Al-Hashemy, S M R Taha, "Voiced-Unvoiced-Silence Classification of Speech Signals Based on Statistical Approaches", Applied Acoustics, pp 169-179, 1988.

[4] C G J Houben, "Automatic Labelling of Speech Using an Acoustic-Phonetic Knowledge Base", ProcEurospeech 1989, pp 104-107 .

[5] N N Bitar, C Y Espy-Wilson, "Knowledge Based Parameters for HMM Speech Recognition ", Proc ICASSP 1996, pp 29-32.

[6] D Yarrington, H T Bunnell, G E Bell, "Robust Automatic Extraction of Diphones with variable Boundaries", Proc Eurospeech 1995, pp 1845-1848.

[7] T.Svendsen, K. Kvale, "Automatic Alignment of Phonetic labels with continuous speech", Proc ICSLP 1990 pp 997-1000.

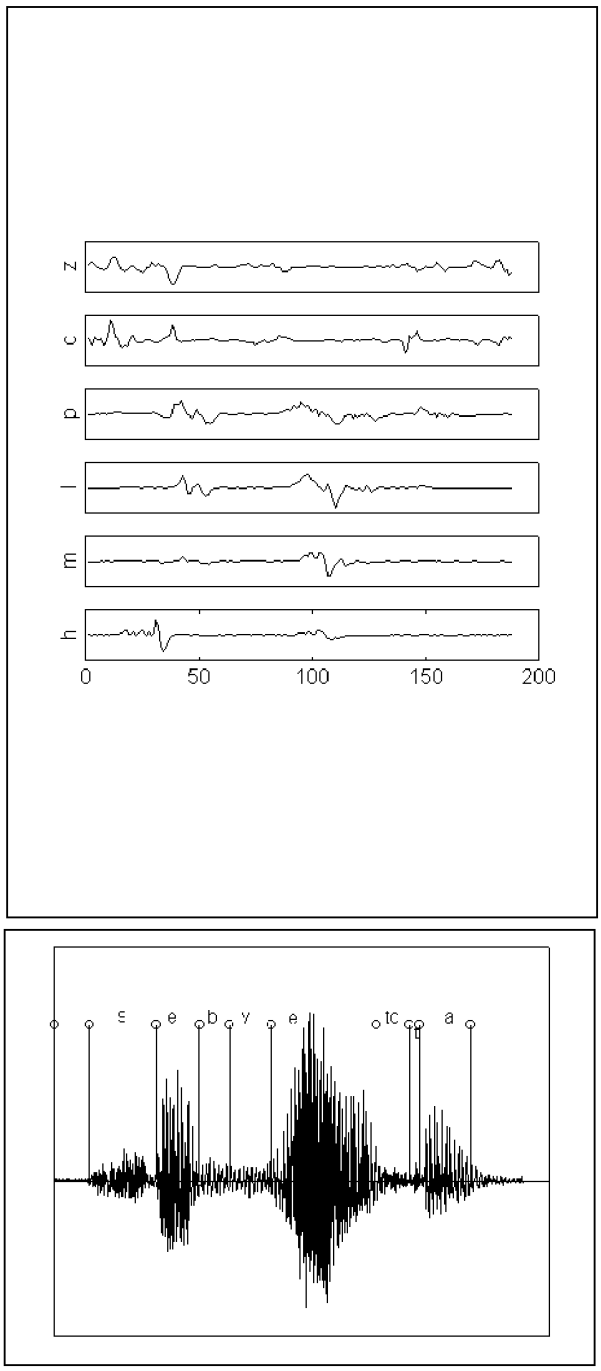

Figure1

Rate parameters from top are zerocrossing rate; correlation rate; energy rate; low frequency, midfrequency and high frequency rate. The features are for dummyword ' sebveta ' .