DYNAMIC FEATURES IN CHILDREN'S VOWELS

Steve Cassidy *Catherine Watson*

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

ABSTRACT

As part of a long term project to develop speech recognitions systems for young computer users, specifically children aged between 6 and 11 years, this paper presents a preliminary investigation into the classification of children's vowels.

In earlier studies of adult speech we found that dynamic or time-varying cues were useful in classifying diphthongal vowels but provided no advantage for monophthongs if duration is included as an additional cue. In this study we investigate whether dynamic cues (modelled by Discrete Cosine Transform coefficients) are present to a greater or lesser extent in children's vowels. Our hypothesis is that some of the observed variability in children's vowels may be due to systematic time-varying features.

We found that the children's monophthong data was better separated by a combination of DCT coefficients and vowel duration than by the formant data sampled at the vowel midpoint plus duration. This result contrasts with our finding on Australian adult data in which we found it was necessary to model the formant trajectory only to separate the diphthongs.

1. INTRODUCTION

While the majority of speech recognition research is based on adult speech, younger computer users, especially those who are learning to read, also stand to benefit from improving speech recognition technology. In developing systems to deal with children's speech it is important that we understand the qualitative and quantitative differences between child and adult speech. This paper presents a preliminary study of the characteristics of children's vowels and, in particular, the presence of dynamic cues to vowel identity in children's speech.

The small number of existing acoustic studies of children's speech[1] point to the increased variability of children's speech compared to adults. Katz, Kripke and Tallal [4] describe increased variability in vowel formant frequencies and voice onset times for consonants. In studies based on acoustic and video data, they found that the increased variability is not accounted for by increased anticipatory coarticulation in young children. Contrary to these results Nittrour *et. al* [6] find that children's fricatives display greater differences in the region of the second formant as a function of the following vowel. They also observe that children showed more distinct spatial targets for vowel gestures than adults, and, in earlier studies, that fricative gestures are less spatially distinct than for adults. This implies that while children's speech production is developing, they may show quite different acoustic patterns to adults.

In earlier studies of adult speech [8, 2, 7] we have looked for evidence of dynamic cues to vowel identity by means of multiple spectral slices and simple models of formant trajectories.

The results of these studies was that this dynamic information only provided additional cues for monophthongs if duration was not included as an additional parameter. In this case the dynamic model is able to capture the different formant trajectories resulting from early or late targets in different vowels [7]. If duration is included, the monophthongs are adequately identified from the formant values at the vowel target. Diphthongs, which have two targets, are better identified when dynamic information is included with duration since the shape of the formant or spectral trajectory is different for each vowel.

In the experiments reported here, we compared classification scores for a single set of formant values taken at the vowel target with those for formant tracks modelled with three Discrete Cosine Transform (DCT) coefficients. As a comparison, a similar experiment was carried out using adult data selected to parallel the vowel contexts of the children's data as closely as possible.

[1]These studies refer to children between the ages of 3 and 10 years

	ɜ	oʊ	æ	ɛ	ɪ	ɒ	ʊ	ʌ
Children	37	48	72	31	25	33	32	56
Adult	54	53	55	52	54	53	53	52
	a	aɪ	aʊ	eɪ	i	ɔ	ɔɪ	u
Children	55	24	24	33	64	34	23	49
Adult	53	52	51	54	52	52	54	52

Table 1: The distribution of vowel tokens in the children's and adult data.

2. MATERIALS

A database of isolated word utterances was collected from 8 children (4 boys and 4 girls) aged between 7 and 11 years. All children were native speakers of Australian English. Each child produced a set of 129 isolated words containing examples of all the monophthongs and diphthongs of Australian English in various phonetic environments. Recordings were made in a sound treated studio onto DAT. The data was segmented and labelled phonetically. The vowel onset and offset and the position of one or two targets were marked for each vowel. The labelling criteria are the same as to those discussed in [3].

The first three vowel formants were automatically tracked using the ESPS formant tracker by Entropic (75ms window, 5ms frame step, 10th order LPC). The formant tracks were hand corrected where they had been mis-tracked.

For this study, lexically stressed vowels (monophthongs and rising diphthongs) were selected from the database which were either preceded or followed by a stop or fricative consonant, or which were in word initial or final position and followed or preceded by a stop or fricative. Formant data was extracted for each vowel at the vowel target, at the vowel midpoint, and for the DCS study, as a continuous track across the whole vowel. Database queries and data extraction were performed with the Emu speech database

system [1].

Adult data was selected from the ANDOSL [5] database to match the children's data as closely as possible. Vowels were selected from the isolated word data of five male speakers which were preceded by fricatives or stops and which received primary lexical stress. All of these vowels were followed by [d]. Since the ANDOSL data was not marked with vowel targets we assume that the vowel target occurs at the midpoint; earlier studies have shown that this assumption can be made with only a small (non-significant) drop in classification accuracy.

The distribution of tokens in the children's and adult data is shown in Table 1.

3. VOWEL DURATION

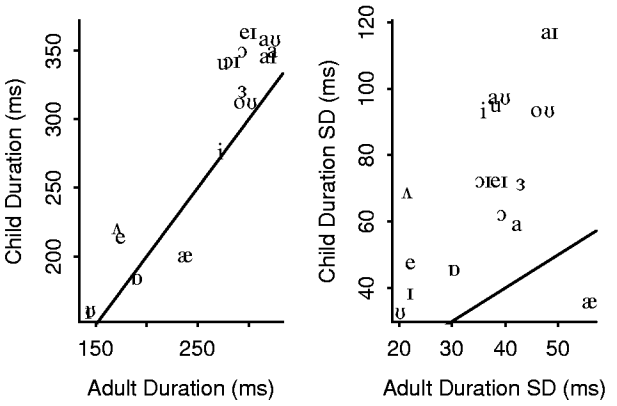

Figure 1: Duration differences for child and adult vowels. The left plot shows the differences in duration, the right plot shows differences in standard deviation. Vowels above the diagonal line are longer or more variable in length for children, those below are shorter or less variable.

A comparison of the mean durations of the child and adult vowels (Figure 1) shows that, in general, the children's vowels were longer than those of the adult males. The second plot in Figure 1 shows that the standard deviation of the vowel durations for children is also much larger than that of the adult male talkers. Although these figures do not relate directly to the classification experiments reported below, they do serve to illustrate the increased variability of the children's data.

4. CLASSIFICATION EXPERIMENTS

All classification experiments were performed using the first two formants sampled at the vowel target or midpoint, or the first two formant trajectories. Initial experiments showed that adding a third formant did not produce a significant increase in classification accuracy. The distribution of the children's monophthong vowel data on the F1/F2 plane is shown in Figure 2. From this plot we can see that we should expect significant confusions between tense/lax vowel pairs ([i]/[ɪ], [a]/[ʌ], [ʊ]/[ɔ]) based on formant data from the vowel target.

To capture dynamic cues to vowel identity, we modelled the vowel formant trajectories with three Discrete Cosine Transform (DCT) coefficients. These three coefficients model the mean, slope and curvature of the formant trajectory and have been found

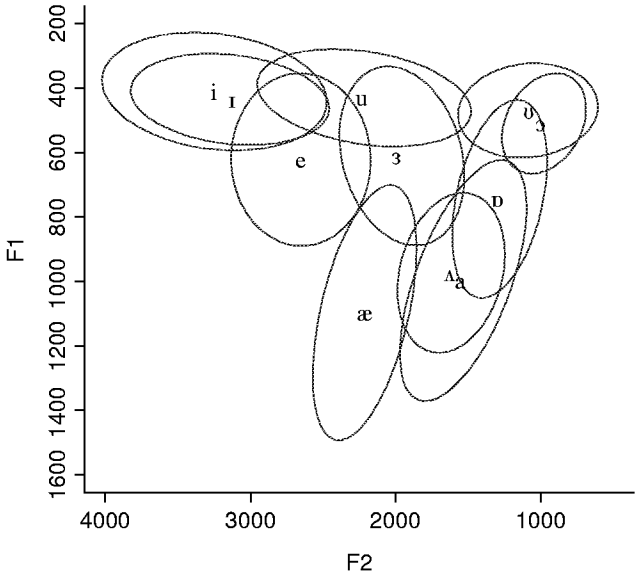

Figure 2: The distribution of the monophthongs from the children's vowel data on the formant plane from data sampled at the vowel target. Ellipses cover around 95% of all data points.

to enhance the separation of vowels in earlier studies [8, 2, 7]. Classification scores using three DCT coefficients for each modelled formant (six features in all for a two formant model) were compared with those for formant values taken at the vowel target or midpoint (two features for a two formant model). In addition, the duration of the vowel token was used as an extra feature in classification.

Classification experiments were done by finding class centroids and covariance matrices for a training set of vowels and then classifying a test set of vowels by measuring the Bayesian distance to each class centroid.

In order to provide sufficient training tokens, all classification experiments were done on a round-robin basis. In each round, all but one speaker were used to train the model (estimate the class centroids) and the remaining speaker was used for testing. Hence all results quoted here are from *open* tests where the training and testing data are different. The results of all training-testing rounds are then summed to give overall scores.

4.1. Results: Adult Data

The experiments with the adult male data reproduces the pattern observed in earlier studies [2, 7]: that the use of dynamic cues (DCT coefficients) is only useful in monophthongs if duration is not included as a separate feature (Table 2). When duration is included, the classification score for the formants at the vowel midpoint (87.29%) is almost identical to that for the DCT coefficients (87.11%). An advantage of using DCT coefficients with duration is only shown when the diphthongs are included in the vowel set.

4.2. Results: Children's Data

An initial comparison between the vowel target and midpoint data shows no significant differences in performance. Since only midpoint data is available for the adults, we will only discuss the results from the midpoint experiments here.

The most interesting observation in the children's results is that

Condition	Children		Adult Male	
	Monophthongs	All Vowels	Monophthongs	All Vowels
Target	61.68	48.43	-	-
Midpoint	63.31	48.59	64.60	45.63
3 DCT	76.02	73.75	76.12	79.08
Target + Duration	76.84	62.03	-	-
Midpoint + Duration	76.84	63.90	87.29	64.78
3 DCT + Duration	84.43	81.88	87.11	84.87

Table 2: Summary results for vowel classification experiments. Scores are percentage correct classifications.

there is a significant ($t = -3.08, p = 0.01$) difference between the overall scores for the midpoint data and the DCT coefficients for monophthongs when duration is included. This is, on the surface, counter to our earlier results and those for the adult data reported above. Closer examination of the results show that the difference is due to a 36% reduction in confusions between [i] and [ɪ], and a smaller 12.5% reduction in confusions between [ʌ] and [a] (confusions of [ʌ] with [ɜ] and [u] are also reduced by a small amount, giving a 21% improvement overall). Both of these differences are significant at the 0.05 level. Differences in the scores for other vowels were small and did not reach significance.

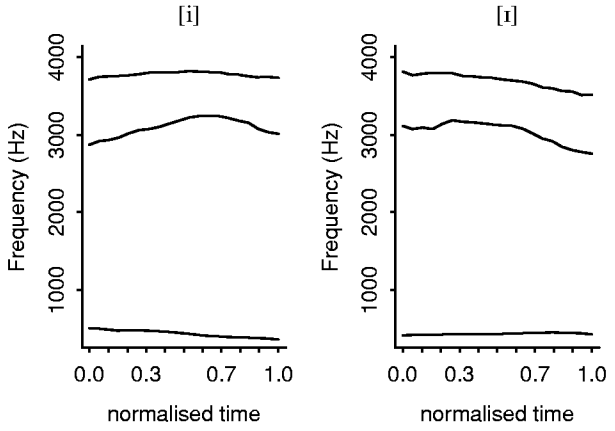

Figure 3: Average formant trajectories for [i] (left) and [ɪ] (right) vowels from the children's data.

The improvement in the classification scores for [i] is explained by the observation that this vowel in Australian English is often diphthongalised [3]. A plot of the averaged formant trajectory for [i] and [ɪ] vowels[2] shows that there is indeed substantial movement in the F2 of [i] which might be consistent with two vowel targets. Whether this vowel is diphthongal or not, it is evident from the plots that these two vowels have very different shaped formant trajectories for F2. This difference is captured by the DCT coefficients and results in the increased classification scores observed.

The results for [ʌ] show a marked reduction in confusions with [a]. If we examine the averaged formant trajectories of these two vowels we see that while the formant trajectories seems similar,

[2]These plots are generated by averaging the formant trajectories of all tokens aligned at the vowel target and truncating the averaged trace to the mean start and end times for all segments relative to the vowel target. The time axis is normalised between 0 and 1.0.

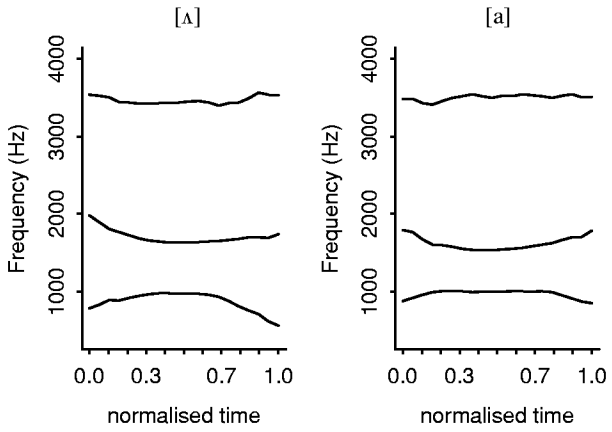

Figure 4: Average formant trajectories for [ʌ] (left) and [a] (right) vowels from the children's data.

there is perhaps more curvature in the first formant of [ʌ] and less in its second formant. These features, encoded via the DCT coefficients, are sufficent to increase the separation of these two vowels. It is interesting to note the large difference in durations for these two vowels (Figure 1), which might have lead to a good separation if it were not for the large variability in the duration of [ʌ].

It is possible that this apparently characteristic formant trajectory might be due to the particular range of consonant contexts from which the [ʌ] tokens were extracted. Of the 56 [ʌ] tokens, the contexts were as follows, with eight (one per speaker) vowels in each context:

pʌb tʌb ʒʌd bʌt gʌt dʌv bʌz

More than half of the following contexts are voiced sounds which may have an effect on the offset of the F1 of the vowel. Further analysis of the results based on following context might reveal whether this is in fact the case. However, given that the adult data consists only of vowels preceding [d] we would expect to see a similar shape in that data. The difference between the two data sets may be the smaller variability in duration in the children's data. The interesting observation here is that this formant trajectory is sufficiently different to that of [a] to increase classification scores when DCTs and duration are used together.

5. DISCUSSION

This study set out to examine some of the acoustic properties of children's speech relative to that of adult talkers. We have found

that there is a general increase in the variability of all parameters (F1, F2, duration), and that some of this variability is systematic and can be used to help in the identification of vowels. The use of a DCT based model of formant trajectories helps to differentiate even the monophthongal vowels in the children's data, contrary to our findings with adult male data. The source of the improvement is a characteristic formant trajectory for some vowels which provides cues additional to those provided by the formant values at the midpoint and the vowel duration.

These experiments suggest that the use of dynamic, time varying features, may help to offset some of the problems associated with added variability in children's speech.

REFERENCES

[1] S. Cassidy and J. Harrington. EMU: an enhanced hierarchical speech data management system. In *Proceedings of the 6th International Conference on Speech Science and Technology*, pages 361–366, Adelaide, 1996.

[2] J. Harrington and S. Cassidy. Dynamic and target theories of vowel classification: evidence from monophthongs and diphthongs in Australian English. *Language and Speech*, 37:357–373, 1994.

[3] J. Harrington, F. Cox, and Z. Evans. An acoustic phonetic study of broad, general, and cultivated vowels. *Australian Journal of Linguistics*, in press.

[4] W. F. Katz, C. Kripke, and P. Tallal. Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual and video data. *Journal of Speech and Hearing Research*, 34:1222–1232, 1991.

[5] J. Millar, J. Vonwiller, J. Harrington, and P. Dermody. The Australian National Database of Spoken Language. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, volume 2, pages 67–100, Adelaide, 1994.

[6] S. Nittrouer, M. Studdert-Kennedy, and S. T. Neely. How children learn to organise their speech gestures: further evidence from fricative-vowel syllables. *Journal of Speech and Hearing Research*, 39:379–389, 1996.

[7] C. Watson and J. M. Harrington. Acoustic evidence for dynamic formant trajectories in australian english vowels. *Journal of the Acoustical Society of America*, In Press.

[8] S. A. Zahorian and A. J. Jagharghi. Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, 94:1966–1982, 1993.