ACOUSTIC COMPARISON OF CHILD AND ADULT FRICATIVES

Akiko Onaka and Catherine I. Watson .
Speech Hearing and Language Research Centre
Macquarie University

ABSTRACT: The study presents acoustic comparisons between child and adult productions of the 9 English fricatives. Fricative tokens are obtained from citation-form words by 4 boys and 4 girls (aged 7 to 11) and 5 men and 5 women. The results show the overall spectral shapes of fricatives produced by the children are similar to those by the adults, however, some significant differences are found in the resonance values and resonant bandwidth. Classification experiments results show that in general the children's fricative data performed much more poorly than the adult data. The implications of the results for automatic speech recognition are discussed.

INTRODUCTION

This study presents acoustic comparisons of fricatives from children and adults and looks at the implications of these results for automatic speech recognition. Existing speech recognition systems have been developed mainly to recognise adult speech, particularly from adult male speakers and accordingly children's speech possesses some problems for current speech recognition systems (Russell *et al.,* 1996). However, few studies have investigated the performance of speech recognition technology with children. A study by Wilpon & Jacobsen (1996) investigated the performance of speech recognition systems on different age groups and found the recognition rates were lowest for children (aged 8-12). Since children's speech production development has been observed to continue until as late as 10 to 12 years of age (e.g. Tingley & Allen, 1975), there will be more variability in children's speech than in adults. Further, fricatives are one of the most difficult class of sounds for English-speaking children to acquire (e.g. Ingram *et al.,* 1980). Therefore, automatic recognition of children's fricatives may be problematic.

Previous studies have described various acoustic characteristics on fricative consonants in adult speech and some earlier descriptions of adult fricatives (e.g. Hughes & Halle, 1956; Tabain, 1998) have shown that distinctive characteristics of fricatives can be found in their spectral shapes. Children's fricatives have not been studied as extensively as adult fricatives in terms of their acoustic features. Pentz *et al.* (1979) found that the children's spectra of /f, v, s, z, S, Z/ had higher major resonances than the adults' spectra. Nittrouer (Nittrouer *et al.,* 1989; Nittrouer, 1995) found that the spectra of two fricatives /S/ and /s/ produced by children were significantly different from those of adult fricatives in both the position and shape of the resonant peak.

The first part of this study aims to examine the spectra of all the 9 English fricatives in children's speech and compare them to those from adults. The comparisons are quantified using spectral moments. In the second part we perform a series of small scale classification experiments to investigate whether there are differences in the fricative separation rates between the adult and child data. We discuss the implication of the results for speech recognition systems for children.

METHOD

Subjects

Children's fricatives analysed in this study were taken from a pre-existing database (Cassidy & Watson, 1998). The speech samples were collected from 4 boys and 4 girls aged between 7 and 11 years. All subjects were native speakers of Australian English (AE). The adult fricatives analysed were taken from the Otago speech database (Sinclair & Watson, 1995). The database included 11 men and 10 women aged between 16 and 33 years as subjects. Among these speakers, 5 male and 5 female speakers were selected for this study. All the subjects were native speakers of New Zealand English (NZE). The Otago database was used because it uses the same speech list materials as the Australian children's database. Ensuring maximum control of the (phonetic) context in which fricatives

were produced prevents us from attributing differences between child and adult data to the effects of neighbouring segments on sounds in question.

Materials

The speech materials used in the two databases consisted of a set of citation-form productions of 129 (real) words from each subject. Each database contained usually three target words for each phoneme, which exemplified the phoneme in word-initial, word-medial, and word-final positions. For this study, 26 words containing an instance of the fricative sounds /f, v, T, D, s, z, S, Z, h/ were selected. Although their phonetic contexts were not controlled, fricatives which were either preceded or followed by [i] vowel were avoided due to a potential coarticulation effect on fricative sounds. Although AE and NZE vowel spaces are very similar, there are some noticeable differences, in particular the /i/ vowel is quite different being a high front vowel in AE and a retracted mid vowel for NZE (Watson *et al.*, 1998).

As described in Cassidy & Watson (1998), the child data was recorded in a sound-treated studio at the Speech Hearing and Language Research Centre, Macquarie University. The speech data was sampled at 20 kHz and quantised to a 16-bit number. The adult speech database was recorded in a quiet room. The speech was sampled at 22.05 kHz and quantised to a 16 bit number (Sinclair & Watson, 1995). For both of the data, the material was presented to the subjects in a random manner, one word at a time to avoid list intonation. One token for each target word was obtained for the child data and three tokens for the adult data.

The children's speech data was segmented and labelled phonetically by a trained phonetician (Cassidy & Watson, 1998) and the adult speech data was labelled phonemically by the first author. The labelling was carried out in EMU, a speech data management system (Cassidy & Harrington, 1996). The labelling criteria followed those described in Croot & Taylor (1995).

Spectra and spectral moments

Spectra for the adult and children's fricative tokens were obtained by performing a series of overlapping 256 points FFTs across the entire fricative token and averaging the result. Each successive FFT slice was overlapped 50 % over the previous FFT slice. The bandwidth of the spectra from the child data was 10 kHz, whereas for the adult data it was 11.025 kHz, due to the slight different sampling frequencies of the two data sets. In Figure 1 the adult spectra were truncated to 10.0734 kHz (i.e., which is the FFT bin closest to 10 kHz) for comparison purposes. The spectral moments were calculated based on the formula given in Forrest *et al.* (1988) and calculated for the spectra of fricatives in the frequency range of 1 to 9 kHz. The region beneath 1 kHz is only expected to yield information about voicing, since we analyse the voiced and voiceless fricatives separately, we felt there was no need to include this information. Secondly, the children's fricative spectra above 9 kHz all showed a rapid drop in amplitude (see Figure 1), suggesting instrumental distortion due to perhaps the anti-aliasing filter. For this reason the spectra above 9 kHz were not examined either.

Classification experiments

The Gaussian classification technique was used to classify a given token, similar to that used by Tabain (1998). A Bayesian distance metric was used to measure the probability of a given token belonging to a particular class of phoneme, and a "round-robin" procedure was to test and train the classifier. For this study, three classification experiments were carried out: firstly on the adult data, secondly on the child data, and finally on both of the adult and child data where the classifier was trained with the adult data and tested on the child data.

RESULTS

Fricative spectra and spectral moments

Figure 1 shows averaged fricative spectra for the 8 child and 10 adult speakers. The overall spectral shapes of the adult and children's fricatives are similar. The spectra for the adult fricatives are similar to other studies (e.g. Tabain, 1998). The spectra of /f, v, T, D/ are relatively flat with a diffuse spread of energy across the spectra, with /v, D/ having a noticeable voiced peak in the frequency region below 1 kHz. For /s, z/, most of the spectral energy is in the 4-9 kHz region for the children and 4-6 kHz region for the adults. The spectral energy for /S, Z/ is in the 2-6 kHz region for the children and 2-3.5 kHz

region for the adults. In addition, their spectral peaks are narrower compared to the alveolar fricatives. The glottal fricative /h/ has a quite skewed spectrum where the most of spectral energy is concentrated below the 3 kHz region for the children and 1.8 kHz region for the adults. For the most part the resonant peaks for children's spectra are broader than the adult's, as can be seen in Figure 1.

The first four spectral moments values of the adults' and children's spectra are given in Table 1. The first spectral moment gives the centre of gravity (i.e., the mean) of the spectra, and has the dimensions Hertz. The centre of gravity of the voiceless spectra is greater than its voiced counterpart for both the child and adult data (as would be expected). For /h, S, Z, s, z/, the further the constriction from the lips, the lower the centre of gravity (as can be seen in Figure 1).

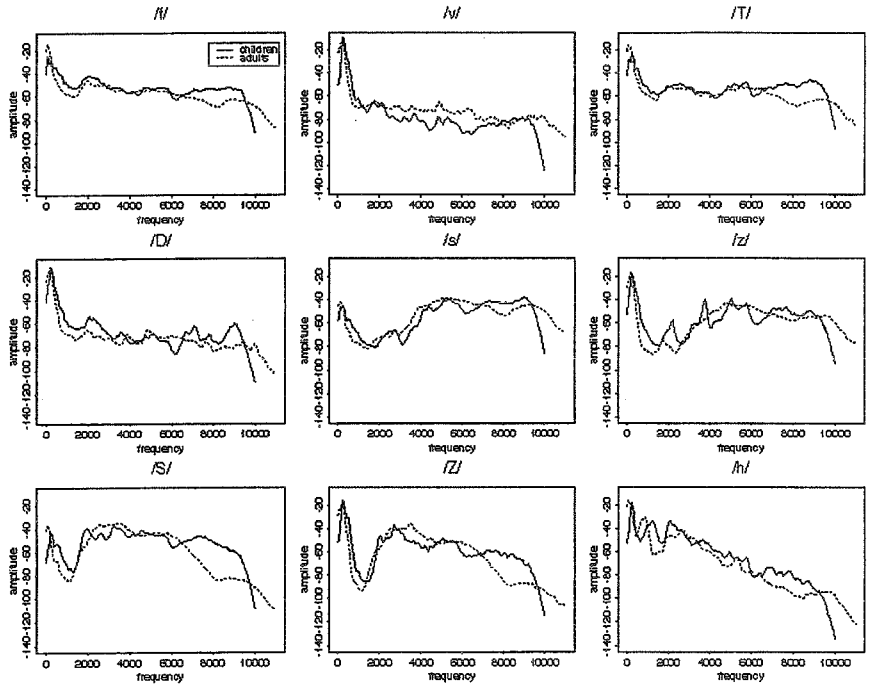

Figure 1. Averaged spectra for the child and adult fricatives. Spectra were averaged across all tokens and subjects for each fricative. The amplitude is presented in 10 log (dB) scale.

Although this trend could be expected to continue for /f, v, T, D/, it does not, but this is probably due to the fact the bandwidth was not large enough to get the spectral peaks for the labio and labio-dental fricatives (cf. Tabain, 1998). For each of the voiceless fricatives except /h/, the centre of gravity of the children's spectra is greater than for the adults, and this difference is significant for /T, s, S/. This is probably due to the fact the children's vocal tracts are smaller than the adults, and hence the resonant frequencies are all higher. The reason why this was not found for the voiced fricative may be due to the fact that not all the voicing component was removed in the children's spectra when truncating the spectra at 1 kHz, due to their higher pitched voices.

The second spectral moment is a measure of the diffuseness of the spectra around the centre of gravity, i.e., the variance. We take the square root of this to give the standard deviation, which has the dimension Hertz. The voice/voiceless pairs show a similar degree of diffuseness for both the adult and child data. The spectra of /f, v, T, D/ for both the adult and child data are more diffuse than for /s, z, S, Z, h/, and this can also be seen in Figure 1. The diffuseness of the children's fricative spectra is always greater than that for the equivalent adult's fricative spectra. This difference is significant for /f, T, D, s, z/. This suggests that the children do not have as finer control over their fricative production as the adults and are not as able to tune their noise as the adult.

	ADULTS				CHILDREN			
	1st (Hz)	2nd (Hz)	3rd	4th	1st (Hz)	2nd (Hz)	3rd	4th
f	3206.3 (709.3)	1944.0 (942.7)	0.56 (0.55)	2.88 (1.14)	3319.2 (1116.1)	2161.0* (1252.7)	0.62 (0.85)	3.11 (1.95)
v	2513.5 (802.6)	1971.1 (1082.7)	1.02 (0.74)	3.85 (2.90)	2298.8 (981.7)	2090.0 (4476.3)	1.52 (1.25)	6.15 (8.06)
T	3678.3 (690.2)	1975.4 (1072.6)	0.13 (0.54)	2.47 (0.82)	4401.8 * (832.2)	2316.1* (1292.3)	-0.20 (0.56)	2.14 (0.69)
D	2663.9 (1016.1)	1951.8 (1113.4)	0.91 (1.00)	4.15 (4.04)	2874.2 (1344.2)	2254.2* (1616.4)	1.11 (1.55)	6.03 (11.16)
s	5116.3 (779.0)	1338.9 (876.7)	- 0.04 (0.80)	4.42 (1.78)	5758.9 * (976.8)	1367.1 (885.3)	-0.64* (1.37)	6.39* (4.12)
z	4921.2 (761.4)	1373.2 (977.3)	0.13 (0.79)	4.60 (2.52)	5274.8 (1001.0)	1531.1 (1150.8)	-0.15 (1.51)	6.23 (7.37)
S	3026.5 (439.6)	1257.6 (698.0)	0.90 (0.56)	3.58 (1.87)	3714.7 * (978.8)	1680.3* (989.8)	0.74 (0.87)	3.58 (2.81)
Z	2915.8 (485.1)	1229.6 (735.9)	1.15 (0.55)	4.26 (2.29)	3460.4 (983.5)	1555.4* (1104.3)	1.30 (1.38)	6.31 (8.06)
h	1782.1 (320.6)	1144.2 (615.5)	1.37 (0.76)	7.29 (4.31)	1715.2 (526.7)	1274.0 (890.7)	1.40 (0.59)	6.57 (3.15)

Table 1. The mean of the first four spectral moments and their standard deviation (in parenthesis) for the adult and children's fricatives. Significant differences are indicated by * ($\alpha < 0.001$); all alpha values were adjusted for multiple testing.

The third spectral moment is a measure of the degree of asymmetry of the spectral distribution about the mean. If it is a positive value, it means the spectrum is skewed to the right of the mean and if it is a negative value, it means the spectrum is skewed to the left of the mean. The third moment is a dimensionless quantity. The spectra of /S, Z, h/ have a positive moment which means the spectra are skewed to the right of the mean, as can be seen in Figure 1. The spectra of /s, z/ are expected to be skewed to the left, and they are except for the adult /z/. If we had a greater spectral bandwidth, the spectra of /f, v, T, D/ would be expected to be skewed to the left too, but at the bandwidth given the spectrum is flat, and hence usually positively skewed. The only significant difference between the adult and child data is for /s/, where the child data is more generally skewed, and this finding was also made by Nittrouer (1995).

The fourth spectral moment is a measure of the peakedness or flatness of the distribution of spectral components relative to a normal curve. If it is flatter, it has a negative value and if it is more peaked, it has a positive value. Like the third moment, it is dimensionless. All the spectra have positive values for the fourth spectral moment. However, it can be seen that the /h/ spectrum is the most peaked, further /s/ is more peaked than /S/, which both in turn are more peaked than /f, T/, and this can also be seen in Figure 1. The values of the fourth spectral moment for the voiced fricatives are all pretty similar. The only significant difference found between the child and adult data is for /s/. The standard deviations of the spectral moments for the child data are usually much greater than for the adults. This is further indication that the production of the children's fricatives is much more variable than the adults.

Classification experiments

Classification experiments were performed on the data using spectral moments (from 1 to 9 kHz), duration and RMS as acoustic parameters for fricative classification. Various sets of these parameters were selected and tested on the data. The best results were from the first four spectral moments and duration, which is what will be reported here[1]. The experiments were carried out on two separate subsets of data, namely a set of voiced fricatives and a set of voiceless fricatives. The results are presented in Table 2.

Adult fricative classification

	f	T	s	S	h		v	D	z	Z
f	**69.2**	26.7	0.8	2.5	0.8	v	**71.1**	25.6	2.2	1.1
T	32.2	**64.4**	1.1	2.2	0.0	D	66.3	**29.2**	4.5	0.0
s	0.8	3.4	**92.4**	3.4	0.0	z	1.1	3.3	**92.2**	3.3
S	6.7	2.2	0.0	**87.6**	3.4	Z	5.1	3.4	1.7	**89.8**
h	0.0	0.0	0.0	6.7	**93.3**					

Children's fricative classification

	f	T	s	S	h		v	D	z	Z
f	**50.0**	31.2	0.0	15.6	3.1	v	**75.0**	25.0	0.0	0.0
T	20.8	**70.8**	4.2	4.2	0.0	D	58.3	**41.7**	0.0	0.0
s	0.0	3.1	**81.2**	15.6	0.0	z	0.0	4.2	**87.5**	12.5
S	20.8	0.0	8.3	**70.8**	0.0	Z	6.2	18.8	6.2	**68.8**
h	18.8	0.0	0.0	0.0	**81.2**					

Children's fricative classification against adult fricatives

	f	T	s	S	h		v	D	z	Z
f	**40.6**	50.0	0.0	3.1	6.2	v	**70.8**	29.2	0.0	0.0
T	16.7	**79.2**	4.2	0.0	0.0	D	70.8	**29.2**	0.0	0.0
s	0.0	3.1	**93.8**	3.1	0.0	z	0.0	8.3	**91.7**	0.0
S	29.2	25.0	33.3	**8.3**	4.2	Z	6.2	12.5	62.5	**18.8**
h	25.0	0.0	0.0	0.0	**75.0**					

Table 2. The confusion matrices generated for the classification experiments

When the classification was done on the adult fricatives using all of the spectral moments and duration as parameters, the voiceless fricatives showed the higher overall classification accuracy than the voiced fricatives (80.5 % and 68.9 %, respectively). Table 2 shows the confusion matrices generated. It can be seen that /h, s, S/ are well classified whereas confusions often occur between /T/ and /f/. Similar confusion patterns to the voiceless fricatives are found in their voiced counterparts. These findings are all expected and are similar to other studies (e.g. Tabain, 1998).

For voiceless fricatives, the overall classification accuracy was lower for the child data than that of the adults (69.5 % and 80.5%, respectively), and for voiced fricatives the accuracy was similar (68.2% and 68.9%, respectively). Table 2 shows, as in the adult data, that the children's labio-dental and dental fricatives, regardless of voicing, are often confused amongst each other. However, other confusions are found in the child data that are not found in the adult's: the children's /f/ was noticeably confused with /S/, as was /s/ with /S/, /S/ with /f/, /h/ with /f/, /Z/ with /D/ and /z/ with /Z/. This suggests that the variability in the children's fricative production (as suggested by the spectral moment results) would complicate the speech recognition process for systems trained with child data.

One possibility to try and improve recognition for children's fricatives is to use adult fricative data to train a recogniser. The classification accuracy using the adult data to train the model, and the child data to test was, however, 59.4% for voiceless fricatives and 55.7% for voiced. The confusion between the labio-dental and dental fricatives, as observed in the previous two experiments remains. However, there is much confusion between other fricative classes, and much more than that observed from training and testing with the child data only. This suggests that using adult data to train a recogniser to be used by children is not a good idea.

CONCLUSION

The results of this study show that although the overall spectral shapes of fricatives produced by the children were similar to those by the adult speakers, some significant differences between the adult and children's fricatives were found in the resonance values and resonant bandwidth. These differences are probably due to the children's smaller vocal tract, and the children's less finer motor control over their fricative production. There was also greater variability in the acoustic measures in the child data than in the adult's. The classification experiments results show that in general the child data performed much more poorly than the adult data. There were much more confusions between fricative classes for the child data. Further work needs to be done to investigate the nature of relevant acoustic differences found between the children and adults (e.g., phonological or coarticulatory in nature) and how these might effect on the classification of children's fricatives.

NOTES

(1) The results for child fricative classification were from the classification using the first three spectral moments and duration rather than the first four spectral moments and duration due to a classification failure with the latter set of parameters.

REFERENCES

Cassidy, S. & Harrington, J. (1996) "EMU: an enhanced hierarchical speech data management system", Proc. of the sixth Australian International Conference on Speech Science and Technology, Adelaide, 361-366.

Cassidy, S. & Watson, C. I. (1998) "Dynamic features in children's vowels", Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), Sydney, 959-962.

Croot, K. & Taylor, B. (1995) "Criteria for Acoustic-Phonetic Segmentation and Word Labelling in the Australian National Database of Spoken Language", URL:http://www.shlrc.mq.edu.au/criteria.html

Forrest, K., Weismer, G., Milenkovic, P. & Cougall, R. N. (1988) "Statistical analysis of word-initial voiceless obstruents: Preliminary data", J. Acoust. Soc. Am. 84 (1), 115-123.

Hughes, G. W. & Halle, M. (1956) "Spectral properties of fricative consonants", J. Acoust. Soc. Am. 28 (2), 303-310.

Ingram, D., Christensen, L., Veach, S. & Webster, B. (1980) "The acquisition of word-initial fricatives and affricates by children between 2 and 6 years", In *Child Phonology* edited by G. H. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (New York: Academic Press), Vol. 1, p.169-192.

Nittrouer, S. (1995) "Children learn separate aspects of speech production at different rates: Evidence from spectral moments", J. Acoust. Soc. Am. 97 (1), 520-530.

Nittrouer, S., Studdert-Kennedy, M. & McGowan, R. S. (1989) "The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults", J. Speech Hear. Res. 32, 120-132.

Pentz, A., Gilbert, H. R. & Zawadzki, P. (1979) "Spectral properties of fricative consonants in children", J. Acoust. Soc. Am. 66 (6),1891-1893.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B. & Baker, P. (1996) "Applications of automatic speech recognition to speech and language development in young children", Proceedings of ICSLP 96', Philadelphia, 176-179.

Sinclair, S. & Watson, C. (1995) "The development of the Otago speech database", Proc. of the 2nd New Zealand International Conference on Artificial Neural Networks and Expert Systems, 298-301.

Tabain, M. (1998) "Non-sibilant fricatives in English: Spectral information above 10 kHz", Phonetica, 55, 107-130.

Tingley, B. M., & Allen, G. D. (1975) "Development of speech timing control in children", Child Development, 46, 186-194.

Watson, C. I., Harrington, J., & Evans, Z. (1998) "An acoustic comparison between New Zealand, and Australian English Vowels", Australian Journal of Linguistics, 18 (2), 185-207.

Wilpon, J. G., & Jacobsen, C. N. (1996) "A study of speech recognition for children and the elderly", Proceedings of ICASSP 96", Philadelphia, 349-352.