Development of Phonological Categories in Children's Perception of Final Voicing in English

CAROLINE JONES
MARCS Auditory Laboratories, University of Western Sydney
c.jones@uws.edu.au

1. Introduction

Research into speech perception over the last fifty years has increasingly tended to support the conclusion that the adults' perception of phonological categories is characterised by the use of multiple acoustic properties. There is typically no one acoustic property which is necessary and sufficient to signal the value of a phonological category, and there is a high degree of redundancy among these acoustic properties as cues. For example, the number of acoustic properties which have been shown to affect the perception of intervocalic voicing is at least sixteen (Lisker 1986). The use of these properties is complicated by the lack of acoustic invariance in speech: the extent of evidence for a category available from any one property in a given dialect or language varies as a function of phonological context, and differs subtly from one utterance to the next, and from speaker to speaker. Adults adjust for all this variability in using multiple acoustic properties to perceive phonological categories. What is the course of development of this ability, in childhood?

Few studies have addressed the question of how children's perceptual phonological categories develop to become adultlike in using multiple acoustic properties. Just two phonological contrasts have been investigated in a systematic, detailed way. These contrasts are: (1) the presence of a prevocalic stop (following a fricative, i.e. "say" vs. "stay" – Morrongiello, Robson & Clifton 1984); (2) syllable-initial, prevocalic, fricative place of articulation (Nittrouer & Studdert-Kennedy 1987). In the Developmental Weighting Shift model (DWS – Nittrouer 2002), the results of these studies are argued to be evidence that children differ from adults in weighting formant transitions more heavily than adults in their perception of phonological categories. We need more evidence, from a wider variety of phonological contrasts, to evaluate this suggestion. As pointed out by Gerrits (2001), for example, the argument that the results of Morrongiello et al. 1984 are evidence for the DWS is quite problematic.

A well-known aspect of adults' speech perception is that the cues to a phonological category are often in a trading relation (Repp 1982), where a decrease in the evidence provided by one cue can be offset by an increase in the evidence provided by another cue. The existence of trading relations in adult speech perception reflects the flexibility which adults have developed to deal with the lack of invariance in speech production. Thus, an alternative way in which children's speech perception may differ from adults is in the flexibility with which children make use of information from various acoustic properties in their phonological decisions. It is this possibility that is investigated in the present paper, through an experimental study of children's and adults' use of vowel duration and first formant (F1) offset in the perception of final stop voicing (/t/ vs. /d/) in Australian English. In the production of postvocalic stop consonants in English, the vowel is typically longer before /d/ than /t/ (House & Fairbanks 1953), and F1 typically falls further during the vowel offset before /d/ than /t/ (Hillenbrand, Ingisano, Smith & Flege 1984). The

vowel duration difference is known to be used by adults in judgments of following consonant voicing in English (Denes 1955 and many subsequent studies), and also by children, perhaps less reliably (Greenlee 1980, Krause 1982). F1 information, particularly during vowel offset, is known to be used by adults (Summers 1988, Crowther & Mann 1992) in the perception of final voicing in English. No studies have systematically investigated children's use of F1 (or in fact any cues other than vowel length in the perception of final voicing). In sum, final voicing offers a good opportunity to study children's use of multiple cues for a phonological contrast for which adults' perception is relatively well understood.

2. Methods

2.1 Stimuli

Fourteen auditory stimuli were presented to listeners for identification as "heart" or "hard". These words were chosen for the task because they are a minimal pair of words which are picturable, would be familiar to children, and are of similar relatively high frequency. The CELEX (version 2.5) wordform log frequency for "heart" is 2.1614 and for "hard" it is 2.2672.

The fourteen stimuli formed two seven-step continua. Each continuum had seven equal steps of vowel length. The two continua differed in the relative fall of F1 during the last 50 ms of the vowel. The exact values for vowel length and F1 offset fall for the fourteen stimuli are shown in Table 1. These values were chosen from the range of natural values in speech, following pilot testing with adults; the 120 Hz vs. 240 Hz F1 offset difference reliably affects category judgments but does not swamp the dominant effect of vowel length, for adult listeners. The F1 offset and vowel length differences were also found in pilot testing with children to be large enough for children to make use of them perceptually. In this way, the investigation of the extent to which children use multiple cues is not prejudiced at the outset by the use of acoustic manipulations which are too small for children to detect and use.

Table 1. Vowel length and F1 offset values of the stimulus set

Stimulus number	Vowel length (ms)	F1 offset fall (Hz)
1	126	120
2	153	120
3	180	120
4	206	120
5	233	120
6	260	120
7	293	120
8	126	240
9	153	240
10	180	240
11	206	240
12	233	240
13	260	240
14	293	240

All stimuli were constructed by recording, resynthesizing, and modifying versions of a natural voice recording. An adult male speaker of Australian English recorded the word "hard" several times onto DAT tape, in the carrier phrase "it's a twenty _ length". One recording was chosen for use. This recording was then downsampled to 10 kHz, and the two different 50ms F1 offsets were generated using the Kay CSL LPC Parameter Manipulation/Synthesis Program (Model 4304, Software version 1.X). These two F1 offsets differed in the frequency values for F1 in the last 50 ms of the vowel. In the higher F1 offset, F1 fell 120 Hz, from 770 Hz linearly to 650 Hz at the end of the vowel. In the lower F1 offset, F1 fell 240 Hz, from 770 Hz linearly to 530 Hz at the end of the vowel. To make the stimulus set shown in Table 1, each of the two vowel offsets was added (using CoolEdit 2000) to seven different various lengths of preceding resynthesized CV sequence ([ha], derived from the recording of "hard" from which the vowel offsets were generated). The waveform adding was done in CoolEdit 2000, and the resulting audio files were free of any clicks or other artifacts resulting from the resynthesis or adding of waveforms. All stimuli were equalized for amplitude. 25 ms of silence was added to the beginning and end of each stimulus.

2.2 Participants

Fifteen adults and seven children participated as listeners in the experiment. The mean age for the adults was 19 years (range 18;3 to 37;6 years). All adult participants were first-year undergraduate students at the University of Western Sydney (Bankstown), who participated for course credit for their introductory psychology course. By chance, all adult participants were female. All adults had self-reported normal hearing at the time of testing, and were self-reportedly free of any history of hearing or speech disorders. They were all native monolingual speakers of Australian English. One additional adult was originally recruited, but this participant's responses were much less consistent than the other adults' responses, perhaps due to inattentiveness; in testing, this listener classified less than 70% of the two endpoint stimuli (Stimuli 1 and 14) in the expected way, despite reaching criterion in training. The data from this listener are therefore not included in the analysis.

The children were recruited from two preschools in Sydney. The children were mostly four-year-olds. This age was chosen because it is within the 3-7 year old age range for which the Developmental Weighting Shift model has been proposed, and because four-year-olds have typically not learned to read and have limited knowledge of sound-letter correspondences. Since we know that learning to read has a lasting effect on speech perception (e.g. Burnham, Tyler & Horlyck 2002) and the interest here is in the natural development of oral language independent of learning to read, four-year-olds are a good age to study. The mean age for the children was 4;7 years (range 4;1 to 5;4 years). Among the children were three girls and four boys. All children were reported by their parents to have normal hearing and speech, no attention problems, and no history of recurrent middle ear infections. All of the children were native monolingual speakers of Australian English. In addition to the seven children whose data are reported here, a further three initially began in the study. The data from these children is not included here, for the same reason as for the adult mentioned above; these three children did not maintain consistency, classifying less than 70% of the endpoint stimuli (Stimuli 1 and 14) in the expected way in the testing phase.

2.3 Procedures

Adults were tested individually in one twenty-five minute session consisting of four blocks of trials. The testing of the children was split into four twenty-minute sessions, which were held on different days, spaced approximately a week apart. All children were also tested individually. Adults were tested in a quiet room in the laboratory; children were tested in a quiet room in their preschool.

The stimuli were presented and the participant's responses recorded via a laptop computer (IBM Thinkpad R31, DirectX Version 8.0) running DMDX Version 2.9.01. (DMDX is developed by Jonathan Forster at the University of Arizona). The stimuli were played over good quality stereo speakers (Cambridge SoundWords PC Works) at approximately 70 dB SPL, and this level was maintained the same across listeners and test days. To maintain the children's interest and enthusiasm, the stimuli were presented in the form of a computer game with friendly cartoon creatures, and they received stickers as rewards and markers of their progress in each session. Adults did exactly the same activity as the children, complete with the cartoon pictures. The responses were entered by the participants using picture-labelled left and right Shift keys on the laptop keyboard.

The testing procedure was as follows. First, four examples of the two stimulus endpoints (Stimuli 1 and 14) were presented, in alternating sequence ("heart" – "hard" – "heart" – "hard"), and participants were encouraged to press the correct key after which they received feedback. After the four examples, the two stimulus endpoints were presented one at a time in random order for identification. Participants received feedback on accuracy during this training. The training continued until the subject got six correct in a row. Then the testing began. Participants did not get any feedback on accuracy during the testing. There were a total of 154 test trials (i.e. 11 repetitions of each of the 14 stimuli). These 154 trials were split into four blocks, which adults completed in one session, the children in four sessions. Block 1 consisted of 28 trials, block 2 of 42 trials, block 3 of 42 trials and block 4 of 42 trials. For adults, there were two endpoint reminder trials between each block, and then the participant moved on to the next block of test trials. Children only completed one block per session, and within this block they received more frequent examples of the endpoint stimuli than adults did; after every 7 or 11 test stimuli in each session, children got a sticker break, and then they did two endpoint reminder trials with feedback before resuming testing.

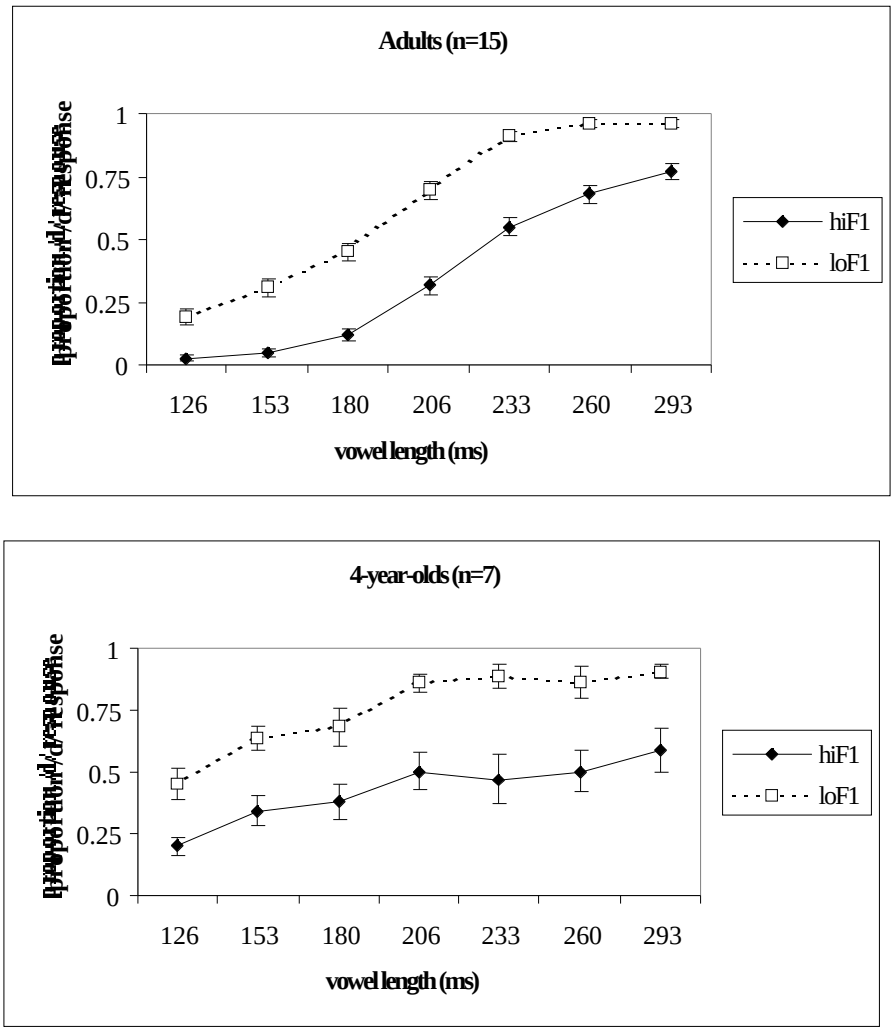
3. Results

Figure 1 shows the proportion /d/ response (with standard error bars) as a function of vowel length and F1 offset value, for pooled data from adult and child listeners. (Note that as yet, fewer children have been tested than adults, which at least partly accounts for the wider standard error bars in the children's data.) Logistic regression analyses were run on the response data, one for the adult data set and one for the children's data set, to assess the effect of vowel length and F1 offset manipulations on listeners' final /t/ and /d/ classification.

The results of logistic regression analysis of the adults' data indicate that the vowel length and F1 offset manipulations each had strong effects on the adults' responses. The best-fitting logistic

regression model for the adult data includes vowel length and F1 as main effects, in addition to a constant. This model is significantly better at predicting the adults' responses than is a constant-only model, $G(2) = 1129.548$, $p<0.001$. The best-fitting model estimates the odds ratios as 2.354 for vowel length, and 6.502 for F1. (The 95% confidence intervals for the odds ratios are (2.196, 2.524) for vowel length, and (5.125, 8.249) for F1.) This means that with every step up in vowel length, the odds of a /d/ response approximately double. And the odds of a /d/ response are about six times greater when the F1 offset is low rather than high. The addition of an interaction term (vowel length by F1 offset) does not improve the predictive power of the model for the adults' data, $G(1) = 0.506$, $p>0.05$. The model is a very satisfactory fit to the data; McFadden's rho-squared is 0.35. (McFadden's rho-squared is a measure of the variability accounted for by the model: Rho-squared = 1- (log likelihood of constant-only model/log likelihood of best-fitting model). Note that rho-squared tends to be much lower than R-squared; values between 0.2 and 0.4 are indicators of very satisfactory fit (Hensher & Johnson 1981).

Figure 1. Adults' and children's identification functions



The best-fitting logistic regression model for the children's data includes a constant, F1 offset as a main effect, and a vowel length by F1 offset interaction. This model has significantly better predictive power than the model which has a constant and main effects but no interaction, $G(1) = 7.952$, $p<0.01$. The model is also an improvement on the constant-only model, $G(3) = 221.65$, $p<0.001$. The model is a fairly good fit to the data, although there is more unaccounted variability for the children's than for the adult data; McFadden's rho-squared is lower, at 0.15. In the model, the estimate of the odds ratio for F1 is 2.280, and the 95% confidence interval for this estimate is (1.262, 4.121).

4. Discussion and conclusion

The results of the analyses indicate that although both adults and four-year-olds in this study made use of both F1 and vowel length properties in judging final voicing, the relative contributions of the two acoustic properties to the phonological categories were different for children as compared with adults.

Adults' classification of the stimuli by phonological category was affected quite strongly by the manipulations of vowel length and F1, and there were compensatory effects of each cue on how adults used the other cue, in their perception of voicing. Thus, for example, at the three shortest vowel lengths (126, 153 and 180 ms), the stimuli containing the lower F1 offset were perceived as voiceless more than half of the time. Similarly, at the three longest vowel lengths (233, 260 and 293 ms), stimuli containing the higher F1 offset were perceived as voiced more than half of the time. The separation between the two F1 functions is least at the endpoint values for vowel length (126 and 293 ms), and greatest at the middle vowel length value (206 ms). This increased separation suggests a classic trading relation between vowel length and F1 offset. When vowel length is at ambiguous values, adults make more use of the information available from the F1 offset than they do when vowel length is more unambiguous.

The results of the analysis of the children's data indicate that in their perception of final voicing for these stimuli, four-year-olds paid particular attention to F1 information, and made use of vowel length information in a more limited way. There is a strong overall effect of F1 on the children's classification, as indicated by the significant main effect for F1 in the best-fitting logistic regression model. The significance of the F1 x vowel length interaction term in that model suggests that the effect of vowel length for children depends on the F1 offset value. When the effect of vowel length is assessed by a separate logistic regression at each of the two levels of F1, this interpretation of the interaction is supported. In the children's data, there is a larger effect of vowel length manipulations when F1 is low than when it is high. The estimate of the odds ratio for vowel length is 1.564 when F1 is low (95% confidence interval = 1.390, 1.759), whereas the odds ratio is 1.268 when F1 is high (95% confidence interval = 1.159, 1.387). Thus, there is an effect of vowel length on the children's classification responses, but the extent of this effect is more limited than for adults, and depends on what value of F1 offset is simultaneously present.

A general conclusion of this study is that the structure of the phonological categories which four-year-olds use in speech perception is similar to, but also interestingly different from the structure of adults' phonological categories. In their perception of final voicing in this study, both adults and children relied on both cues, vowel length and F1 offset, in making their identification

decisions. This finding is consistent with the large literature on the usefulness of vowel length in adults' perception of final voicing (e.g. Denes 1995, and many subsequent studies). The finding that children make use of vowel length is also consistent with previous research (Greenlee 1980, Krause 1982) that tended to find that children aged three years and older can make use of vowel length perceptually, if the acoustic differences are large enough. The adults' response to F1 offset in this study is similar to what was found in the study of vowel length and F1 offset in American English by Crowther & Mann 1992. The use of F1 in the perception of final voicing by children has not been systematically investigated before, in the main because few studies have investigated the use of multiple cues by child listeners, and because vowel length was considered for a time to be the necessary and sufficient cue to final voicing in English.

The results of this study indicate that adults and children differ in the structure of their phonological categories in terms of the extent to which they use acoustic cues in a compensatory way. The adults in this study display a kind of classic trading relation effect, where less evidence from one cue for a given category can be compensated for by more evidence from the other cue for that category. Thus, adults judge give more /t/ judgments to stimuli containing a low F1 offset as vowel length is shortened. And they give more /d/ judgments to stimuli with a high F1 offset when vowel length is increased. Four-year-olds do not do this to anything like the same extent as adults. There also appears to be an asymmetry in the effect of vowel length manipulations on the children's voicing judgments. For the four-year-olds, shorter vowel lengths result in an increase in /t/ responses when F1 offset is low, but there is less of an effect of vowel length manipulations when F1 offset is at the higher value.

The finding from this study that children display smaller compensatory effects among acoustic cues to phonological categories is consistent with the small amount of other research into how children's use of multiple acoustic cues is different from adults'. For example, the perception of five-year-olds shows smaller compensatory effects among F1 onset values and silent gap durations as cues to the presence of a stop in a "say" vs. "stay" judgment (Morrongiello et al. 1984). The way in which the children in the present study make relatively greater use of F1 offset information, and relative less use of vowel duration, can also be compared to children's greater use of formant transitions in perceiving fricative place of articulation (Nittrouer & Studdert-Kennedy 1987, Nittrouer 2002).

In conclusion, the results of this study suggest that four-year-olds' use of multiple acoustic cues in perceiving phonological categories is not characterized by compensatory effects among the cues to the same extent as we find with adults. Why is this the case? One possibility, predicted by auditorist theories of speech perception (e.g. Diehl & Kluender 1989), is that the compensatory effects among cues are smaller for children because the acoustic cues are not integrated to the same extent by children as by adults. Discrimination testing using a subset of the stimuli from the present study is underway to test this hypothesis and find out more about how four-year-olds' perceptual phonological categories differ from those of adults.

References

Burnham D, M Tyler & S Horlyck 2002 'Periods of speech perception development and their vestiges in adulthood' in P Burmeister, T Piske & A Rohde (eds) *An Integrated View of Language Development: Papers in Honor of Henning Wode* Wissenschaftlicher Verlag Trier 281-300.

Crowther C & V Mann 1992 'Native language factors affecting use of vocalic cues to final consonant voicing in English' *Journal of the Acoustical Society of America* 92: 711-22.

Denes P 1955 'Effect of duration on the perception of *voicing' Journal of the Acoustical Society of America* 27: 761-64.

Diehl R & K Kluender 1989 'On the objects of speech perception' *Ecological Psychology* 1: 121-44.

Gerrits E 2001 *The categorization of speech sounds by adults and children: A study of the categorical perception hypothesis and the developmental weighting of acoustic speech cues.* PhD thesis, University of Utrecht.

Greenlee M 1980 'Learning the phonetic cues to the voiced-voiceless distinction: A comparison of child and adult speech perception' *Journal of Child Language* 7: 459-68.

Hensher, D, & L Johnson 1981 *Applied discrete choice modelling* Croom Helm London

Hillenbrand J, D Ingisano, B Smith & J Flege 1984 'Perception of the voiced-voiceless contrast in syllable-final stops' *Journal of the Acoustical Society of America* 76: 18-26.

House A & G Fairbanks 1953 'The influence of consonantal environment upon the secondary acoustical characteristics of vowels' *Journal of the Acoustical Society of America* 25: 105-13.

Krause S 1982 'Vowel duration as a perceptual cue to postvocalic consonant voicing in young children and adults' *Journal of the Acoustical Society of America* 71: 990-95.

Lisker L 1986 '"Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees' *Language and Speech* 29: 3-11.

Morrongiello B, R Robson & C Best 1984 'Trading relations in the perception of speech by 5-year-old children' *Journal of Experimental Child Psychology* 37: 231-50.

Nittrouer S 2002 'Learning to perceive speech: How fricative perception changes, and how it stays the same' *Journal of the Acoustical Society of America* 112: 711-19.

Nittrouer S & M Studdert-Kennedy 1987 'The role of coarticulatory effects in the perception of fricatives by children and adults' *Journal of Speech and Hearing Research* 30: 319-29.

Repp B 1982 'Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception' *Psychological Bulletin* 92: 81-110.

Summers W Van 1988 'F1 structure provides information for final-consonant voicing' *Journal of the Acoustical Society of America* 84: 485-92.