Feature-Cue-Based Processing of Speech: A Developmental Perspective

Stefanie Shattuck-Hufnagel, Helen M. Hanson**, and Sherry Y. Zhao**

*Research Laboratory of Electronics, MIT, 77 Massachusetts Ave., Cambridge, MA 02139, USA
**ECE Dept., Union College, 807 Union St., Schenectady, NY 12308, USA
sshuf@mit.edu, hansonh@union.edu, syzhao@alum.mit.edu

ABSTRACT

Over the past decades a number of research findings have illustrated the extraordinary robustness and flexibility of human speech perception, which combines sensitivity to surprisingly detailed aspects of systematic context-governed variability in word forms with an ability to extract information about the speaker's intended words from minimal information in a sometimes highly reduced signal. Stevens' [19] proposed model of human speech perception, based on the extraction of individual cues to distinctive features, provides an account of this robust perceptual processing in adults, and is also consistent with two recent findings about speech production during development in children learning American English: the occurrence of non-adult-like cues to the voicing contrast in coda stops, and the occurrence of adult-like cues in stop-like productions of voiced dental fricatives.

Keywords: Distinctive feature cues, stop voicing, phonological development, vowel final noise, preaspiration.

1. INTRODUCTION

As speech research has shifted from a focus on individual speech sounds, words and isolated sentences read in the laboratory, to the study of naturally occurring conversational or task-directed speech produced in a communicative context, a number of phenomena have been revealed which pose a challenge to the segment-based approach to speech perception and speech production. These phenomena are instead compatible with the hypothesis that speakers and listeners represent individual cues to the distinctive feature contrasts of their language, and map the feature cue representations onto phonemic lexical representations on the one hand, and onto quantitative signal parameters on the other. Thus, the individual feature cue has a striking advantage as a unit of representation, in that on the one hand it connects directly to the abstract feature, yet on the other hand is able to take on quantitative values related to the speech signal. This idea has been prefigured in the speech literature in ideas like phonetic knowledge [8], phonetic residue [9], cue robustness [20] and cue trading [15] etc.

The older idea that listeners identify successive temporal segments of the speech signal that correspond, to a greater or lesser extent, to the contrastive phonemic categories (or the position-governed allophonic categories) of the language is challenged by many well-known characteristics of the speech signal, such as (1) the **temporal distribution of feature cues** to a given phonemic segment across a relatively broad swath of the utterance, as when the cues to the voicing feature of a coda stop are spread from the longer duration of the preceding vowel to the insertion of an epenthetic vowel after the release; (2) **massive reductions** [7] of acoustic-phonetic information for a word or sequence of words, as when *I'm going to* is produced as something like [amənə], leaving behind a 'phonetic core' of cues to individual features [12]); and (3) the **combination of cues to several allophonic categories** in a single temporal region of the signal, as when a coda /t/ is produced with irregular pitch periods (consistent with a partial constriction formed at the vocal folds) as well as a stop burst (consistent with an oral closure with subsequent pressure buildup and release).

Such observations are difficult to reconcile with the view that adult listeners identify successive acoustic segments in the signal of an input utterance, or that adult speakers create a sequence of successive segments in the signal. Instead, they suggest a model in which speakers and listeners represent and manipulate individual cues to distinctive feature contrasts, that they select context-appropriate cues and compute the parameter values of these individual cues in production, and that they detect both the cues and their parameter values in perception. Additional aspects of speech behaviour also support this view, including **conversational convergence in cue values** (i.e., acoustic-phonetic convergence between two speakers in a conversation, [13]), **cue trading strategies** (e.g. phrase-final lengthening signalled by either a longer steady-state vowel or a slower movement to the following consonant constriction, [3]), and **cue substitution** (in circumstances such as whispered speech in tone languages, where F0 cues can be mirrored in amplitude profiles [4], or prosodic signalling by speakers with motor disabilities, e.g. the substitution of duration for F0 marking of questions by dysarthric speakers [14]. Such observations are compatible with the hypothesis that words are represented as sequences of feature-defined phonemes in the mental lexicon, but that the ways in which speakers and listeners map between the phonological representations in the lexicon and the quantitative aspects of the acoustic speech signal involve the representation and manipulation of individual feature cues and their parameter values.

The feature-cue-based approach to speech analysis raises interesting questions about the development of speech perception and speech production capacities in children. One phenomenon in development that is consistent with the view that speech processing involves the representation of individual feature cues and their parameter values is **covert contrast**. In 1978, Macken and Barton [10] reported that at least some children go through a stage at which they produced distinct distributions for VOT for [+voiced] vs. [-voiced] stops; however, because the values for both of these distributions were within the range for [+voiced] stops in adult productions, this distinction was difficult for adult listeners to perceive. More recent studies have shown that covert contrast is widespread in child speech [5], [17]), and it has even been suggested that most cases of apparent segmental substitution in child speech are illusory---resulting from the use of covertly contrasting cue parameter values by the child [16]. Interestingly, this view suggests that children may have an understanding of the contrastive phonemic categories of their language well before they have control of adult-like cue production. In any case, these findings demonstrate that children's word productions often differ from the adult models they hear around them, and these differences suggest that children (like adults) are able to represent and manipulate individual feature cues and their parameter values.

As in the VOT example from [10], covert contrast in children has generally been discussed in terms of a different distribution of parameter values on the same cues that adults use. However, it is also possible that children use a different cue, with the same consequence: that an adult listener can't intepret this new cue as evidence for the phonological contrast that the child intends. In fact, there are a number of ways that a child's cue selection and cue parameter values may relate to the adult speech of his/her community: the child may use the same cues and parameter value distributions as adults, or may produce different cues and cue parameter values. In this paper we discuss several examples from recent work which illustrate possibilities.

2. EXPERIMENTS

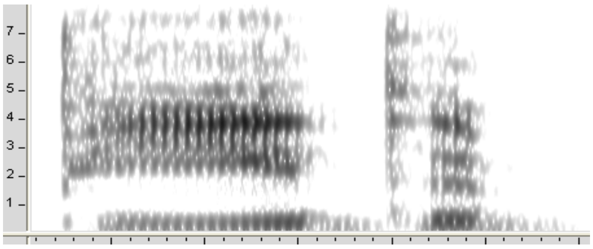
2.1 Corpus

These experiments used data drawn from the Imbrie Corpus [6]. Recordings were made from 10 children and the primary female caregiver of each child. The children were recorded approximately monthly over a six-month period, while the caregivers were recorded just once. The age of the children at the first recording ranged from 2;6 to 3;3. Twenty target words were elicited multiple times from each child during play sessions in which the experimenter prompted the child using pictures and objects. (Further details about the subjects and the recordings can be found in [6], available online.)

2.2 Stop-like /ð/

The phoneme /ð/ is frequently produced in American English, and is often modified to be stop-like in adult speech (Fig. 1).

Fig. 1. Stop-like /ð/ in the utterance *be the (best)*



Nonetheless, listeners distinguish the phonemic contrast between /ð/ and /d/. Zhao ([21], [22]) found that stop-like /ð/ in adult speech frequently occurs when preceded by silence or by stop consonants, occasionally occurs when preceded by fricatives, and rarely occurs when preceded by sonorants and vowels. In the latter contexts it is more likely to be implemented in its canonical form. Spectral analysis ([2], [21]) revealed the following characteristics for stop-like /ð/ compared to /d/:

- Higher burst spectrum peak
- Lower burst spectrum amplitude
- Lower F2 at the following vowel onset

- Higher center of gravity and standard deviation
- Lower skewness and kurtosis

Therefore, stop-like /ð/ in adults differs from /d/ acoustically, in a manner suggesting the preservation of cues to the place feature [dental].

In children's speech, stop-like /ð/ is often cited as an example of a stopping disorder (e.g. [11]). However, based on the adult data cited above, one can ask if it is actually a disorder. That is, are children substituting /d/ for /ð/, changing both place and manner, or changing the manner of production only? Three research questions were posed:

- Does stop-like /ð/ in children's speech occur in segmental contexts consistent with adults?
- Is there acoustic evidence suggesting a difference between children's stop-like /ð/ and /d/?
- Are the children's acoustic data consistent with previous findings from adult data?

The onset obstruent of tokens of *that* and *Daddy* from the first playsession of the Imbrie Corpus were examined.

Context. First, tokens of /ð/ from the word *that* were labeled as being stop-like or not. To be considered stop-like, a /ð/ token needed to exhibit acoustic evidence for (1) formation of complete closure, i.e., a period of no acoustic output at all or a voice-bar decreasing in amplitude, and (2) release of the closure, featuring a sudden onset of acoustic energy. Of 171 /ð/ tokens, 101 (59%) were found to be stop-like. The contexts in which these were found are summarized in Table 1, showing that stop-like productions of /ð/ in children are distributed very like those in adults.

Table 1. Percent stop-like /ð/ in various contexts.

Preceded by:	% Stop-like
Silence	77
Stop	86
Fricative	29
Sonorant	13

Acoustic characteristics. Five of the children produced five or more tokens of stop-like /ð/, and acoustic analysis was limited to the productions of these 5 speakers (99 /ð/ and 49 /d/). Spectral analysis was performed as described in [2] and [22] (see those papers for details of the measures). If stop-like /ð/ differs only from /ð/ in manner and not in place, the following would be expected for /ð/ productions relative to /d/:

- Burst-peak frequency would be higher
- F2 at vowel onset would be lower
- Amplitude of burst spectrum would be lower
- Spectral moments: mean would be higher and standard deviation would be larger
- Spectral moments: Skewness and kurtosis would be smaller

(Again see above papers for the bases of these hypotheses.)

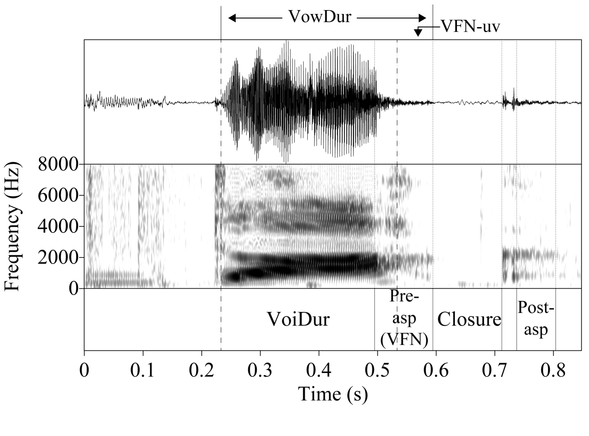
Stop-like /ð/ was found to be significantly different from /d/ in burst-peak frequency, F2 at vowel onset, kurtosis, and skewness. However, unlike adults [21], no significant difference was found for normalized amplitude, spectral mean frequency, and spectral standard deviation. Possible explanations for the lack of significance for those acoustic measures are that (1) the child speech is still developing, and (2) consistent prosodic contexts of tokens are needed for comparison.

In summary, acoustic evidence suggests that children may not be merely substituting /d/ for /ð/. The phonemic and positional contexts in which stop-like /ð/ occur are very similar for children and adults. Likewise, as for adults, child productions have acoustic characteristics suggesting that the place feature of /ð/ is retained, despite the change in manner to non-continuant.

2.3 Vowel-final noise (Preaspiration)

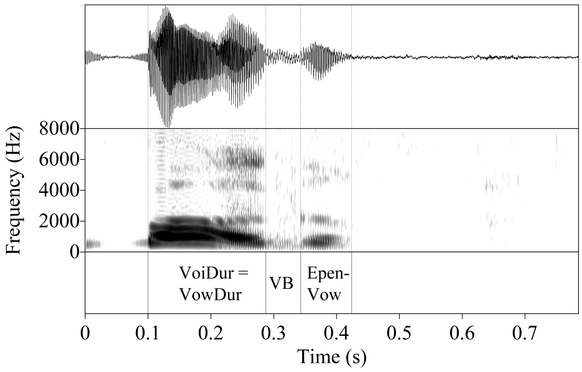
Voiceless stops are not known to be preaspirated in American English, yet preaspirated stops have been reported in child productions (e.g. [18]). Figure 2 shows an example of a preaspirated voiceless coda produced by a child speaker from the Imbrie Corpus.

Fig. 2. Subject C09: *duck* produced with preaspiration (labeled *Pre-asp (VFN)*) and heavy post-aspiration (*Post-asp*). Extracted from the utterance <u>That's</u> not a <u>duck</u>. [1]



Because preaspiration is not necessarily aspiration in fact, we refer to it as vowel-final noise (VFN). Fig. 3 shows an example of a voiced coda produced by a child.

Fig. 3. Subject C01: *bug* produced with a voice bar (labeled VB) and an epenthetic vowel (EpenVow). Extracted from utterance *And ... and Mr. Bug*. [1]

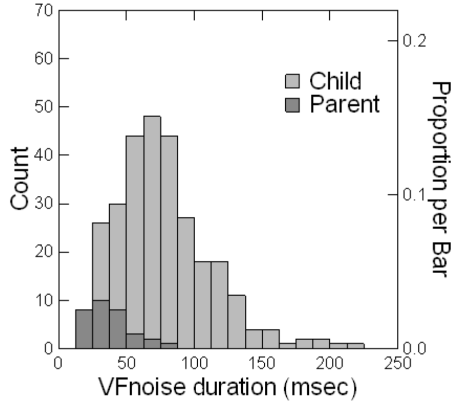


In a study of nine of the ten children and their adult female caregivers in the Imbrie Corpus [6], 1244 tokens of the words *cup*, *tub*, *duck*, and *bug* were analysed to determine the acoustic characteristics of the coda stops [1]. In addition to measuring duration of the vowel-final noise (VFN), vowel duration (VowDur in Fig. 1) was measured as an indication of the child's knowledge of coda voicing. Other acoustic correlates of voicing were also measured, including closure duration (Closure), incidence of voice bars, and duration of voice bars.

Figure 4 shows the distribution of VFN durations. Children were more than twice as likely as adults to produce VFN (29% v. 12%). Both adults and children primarily produced VFN in the voiceless coda context. Average duration of VFN was 78 ms for children, but only 39 ms for adults. Even after normalization (by vowel duration), this difference was significant. Furthermore, children were more likely than adults to devoice vowel endings, and for longer durations, than adults. The conclusion was that children were much more likely than adults to spread their vocal folds early relative to supraglottal closure than were adults.

This result raises the question of whether children produce VFN because of poor coordination of vocal fold spreading and supraglottal closure formation, or whether they intentionally spread their vocal folds early to produce VFN. To answer this, we turned to our data on voice bars. Adults were significantly more likely to produce voice bars than were children (62% v. 42%) for voiced coda stops. Maintaining voicing throughout a stop closure is difficult and it is not surprising if children are not able to do it as frequently as adults.

Fig. 4. Distribution of VFN durations for children and adults. [1]



A possible interpretation of the VFN data is related to the observation that children have difficulty producing voice bars in voiced coda stops. In the coda context, voiceless stops may not be post-aspirated, so that the lack of a voice bar in a voiced stop might jeopardize the signalling of the voicing contrast. As a result, children may spread their vocal folds early for the voiceless tokens, resulting in VFN which strengthens the contrast. The intention may be to produce a cue which provides a contrast for voicing of coda stops, but because adults will likely not perceive or interpret such a cue as the child intends it, it is a *covert cue* (but NOT a covert contrast).

3. CONCLUSION

The feature-cue-based model of speech processing was developed as a model of adult processing, but predicts that children who are still learning to talk represent individual cues to the distinctive features of the phonemes that define and distinguish the words of their language also represent and manipulate individual cues to those features. Several lines of investigation, focussed on analyses of productions by 2-to-3-year-old children in the Imbrie Corpus [6], suggest that children can employ cues and cue parameter values that are different from those produced by adults in the same speech community. These observations are consistent with the hypothesis that language learners can select individual cues and cue parameter values that are suited to their developing capacities. Further studies will be necessary to determine the set of factors that governs these patterns in child speakers.

4. REFERENCES

[1] Anonymous. Vowel final noise and other acoustic cues to laryngeal and supralaryngeal gesture timing in child coda stops. In revision.

[2] Cao, A.Y. (2002). Analysis of acoustic cues for identifying the consonant /ð/ in continuous speech. Master's dissertation, Massachusetts Institute of Technology.

[3] Edwards, J., Beckman, M E and Fletcher, J. (1991). The articulatory kinematics of final lengthening. Journal of the Acoustical Society of America 89, 369-382.

[4] Gao, M. (2002). Tones in Whispered Chinese: Articulatory Features and Perceptual Cues. Thesis, University of Victoria.

[5] Gibbon, F. (1999). Undifferentiated Lingual Gestures in Children With Articulation/Phonological Disorders. Journal of Speech, Language, and Hearing Research 42, 382-397.

[6] Imbrie, A. (2005). Acoustical study of the development of stop consonants in children. MIT PhD thesis.

[7] Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.) Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium. Tokyo, Japan: The National International Institute for Japanese Language. pp. 29-54.

[8] Kingston, J. and Diehl, R. (1994). Phonetic Knowledge. Language 70, 419-454.

[9] Kohler, K. 2000. Investigating unscripted speech: implications for phonetics and phonology. In: Festschrift for Björn Lindblom. Phonetica 57, 85-94.

[10] Macken, M. & Barton, D. (1978). The acquisition of the voicing contrast in Spanish: a phonetic and phonological study of word-initial stop consonants. Journal of Child Language 7, 433-458.

[11] Marshall, C. R., and Chiat, S. (2003). A foot domain account of prosodically-conditioned substitutions. Clinical Linguistics and Phonetics 17, 645–657.

[12] Niebuhr, O. and Kohler, K. (2011). Perception of phonetic detail in the identification of highly reduced words. Journal of Phonetics 39, 319-329.

[13] Pardo, J. S. (2006). On phonetic convergence during conversational interaction. The Journal of the Acoustical Society of America 119, 2382-2393.

[14] Patel, R. (2004). Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast. Journal of Speech, Language, and Hearing Research 45, 858–870.

[15] Repp, B. (1983), Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Communication 2, 341-361.

[16] Richtsmeier, P. (2010). Child phoneme errors are not substitutions. Toronto Working Papers in Linguistics (TWPL) 33, 1-15.

[17] Scobbie, J.M., Gibbon, F., Hardcastle, W.J. & Fletcher. P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In Michael Broe and Janet Pierrehumbert (eds.) Papers in Laboratory Phonology V: Language Acquisition and the Lexicon, 194-207. Cambridge: Cambridge University Press.

[18] Shattuck-Hufnagel, S., Demuth, K., Hanson, H.M. and Stevens, K.N. (2011), Acoustic cues to stop-coda voicing contrasts in the speech of 2-3-year-olds learning American English. In Clements, G. Nick and Rachid Ridouane (eds.), Where Do Phonological Features Come From?: Cognitive, physical and developmental bases of distinctive speech categories. Amsterdam: John Benjamins, 327–342.

[19] Stevens, Kenneth N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. Journal of the Acoustical Society of America 111, 1872-1891.

[20] Wright, R. (2004). A review of perceptual cues and cue robustness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), Phonetically based phonology (pp. 34-57). Cambridge; New York: Cambridge University Press.

[21] Zhao, S.Y. (2007), The stop-like modification of /ð/: A study in the analysis and handling of speech variation. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

[22] Zhao, S. Y. (2010). Stop-like modification of the dental fricative /ð/: An acoustic analysis. Journal of Acoustical Society of America 128, 2009-2020.