SPOKEN LANGUAGE VARIABILITY: IMPLICATIONS FOR MODELING SPEECH PERCEPTION

Keith Johnson
Department of Linguistics & Center for Cognitive Science
Ohio State University

ABSTRACT

Examples of spoken language variability, drawn from a database of conversational speech, illustrate the range of variation that listeners are attuned to. The argument of the paper is that variability in spoken language is neither random nor mechanical but "meaningful".

1. A LINEAR ASSOCIATOR

The nature of variability in spoken language is of particular concern for speech perception theories. One reason that this is important is because many kinds of pattern classifiers assume that deviations from the prototypical pattern for a category are random.

Consider, for example, a simple linear associator. The connectivity matrix (*A*) in a simple linear associator is given by the outer product in (1) below. The vector *g* is the pattern produced by the associator and the vectors *f(i)* are examples of the category. (*f(i)T* is the transpose of the input vector.)

(1) A = sum(g * f(i)T)
(2) f(i) = p + d(i)
(3) A = n*sum(g * pT)

If the category examples *f(i)* have a common prototype *p* and differ from each other by small random deviations *d(i)* as in (2), then the connectivity matrix is a function of the prototype, as in (3). This illuminates a definitional property of a prototype: it can substitute for a collection of examples because it is representative of them.

This example illustrates that the nature of spoken language variability determines whether we can model speech perception using a prototype model, or whether an exemplar-based model is needed. The appropriateness of a prototype model in this illustration hinges on the nature of the variability that the classifier must handle. If the *d(i)* are randomly distributed then listing particular exemplars as in (1) adds nothing to the classifier's performance.

2. A NOTE ON THE SOURCE OF VARIATION

Goldstein ([3]) delineated a problem in characterizing variability in speech production. His study highlights the nonlinearity of the articulation-to-acoustics mapping. In this study, "prototypical" vocal tract shapes in an articulatory synthesizer were perturbed. (They were prototypical in the sense that the mouth shapes were typical, as were the formant values, for carefully articulated versions of these vowels.) The tongue body location parameter of the articulatory synthesizer was perturbed by 2 mm in a circle in the saggital plane. Goldstein's question was whether the acoustic consequences of this articulatory "noise" would be random.

The results for the six vowels tested are shown in figure 1. As can be seen in the variously shaped enclosed regions in this figure the acoustic consequences of uniform articulatory perturbations are nonlinear. For example, the formants of [Q] were hardly affected at all by articulatory variation that shifted [A] over a 100 Hz range in F1 and a 300 Hz range in F2.

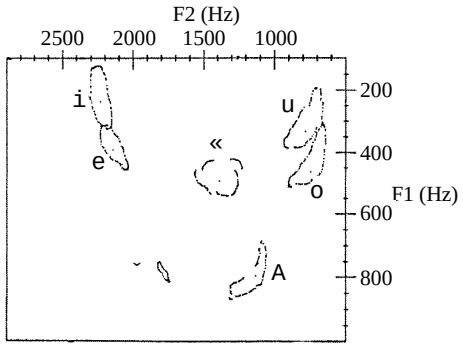

Figure 1. Acoustic consequences of articulatory perturbation. (Adapted from Goldstein, [3])

From this result we can conclude (at least) that if we wish to assume that acoustic variation is random in an acoustic space we must assume that variability in speech production has a complex nonrandom distribution.

3. LAWFUL VARIABILITY

"The major problem that the lack of invariance presents is not that the variation is random or represents noise. In most cases, the causes for the variation are well-understood. Most importantly, they can often be predicted" (Elman & McClelland, [2]).

Elman & McClelland rightly concluded that variability in spoken language is fundamentally nonrandom -- the examples discussed later in this paper reinforce this conclusion. To them variability in spoken language is interesting because it is "lawful". What is meant by "lawful variation" is that variability in acoustic and/or articulatory realizations of speech categories is predictable from context.

For example, Elman & McClelland demonstrated that contextual variability can be exploited in their interactive-activation model of speech perception by allowing vowel phoneme nodes to modify the weights between feature nodes and stop consonant nodes so that context sensitive formant transition cues for stop place of articulation can be appropriately interpreted.

This approach, allowing categorical information to modulate the signaling value of contextual acoustic cues, is probably a good idea because consonant formant transitions really are substantially different in different vowel contexts and this variation is used by listeners in the perception of stop place of articulation. Thus, to exploit the lawful variability of the speech signal, Elman & McClelland's neural network has a somewhat more complicated structure than in a simple linear associator.

Note that a "contextually tuned" model such as Elman & McClelland's network does not have invariant prototype representations of phonemes. For example, because vowel activation can alter the weights, there are as many mappings between [acute] and /g/ as there are states of vowel activation. In this sense, their model uses a multi-prototype representation of /g/.

Another type of contextual variability, retroflexion, that fits nicely with Elman & McClelland's approach can be drawn from the "Variation in Conversation" (vic.psy.ohio-state.edu) database that we are developing at Ohio State. The database consists of a sample of 40 Ohio talkers, stratified for age and sex. Talkers were told that

the research team is interested in how people express their opinions in conversation. A modified sociolinguistic interview format was used to elicit a conversation about everyday topics such as politics, sports, traffic, schools. Talkers wore a head-mounted microphone and conversations were recorded on a small portable DAT recorder. This procedure produced about 40 hours of high quality recordings of completely unmonitored speech.

In this database we have numerous examples of coarticulatory retroflexion. For example, words beginning with /tr/, often start with an affricate [tS] instead of a plain [t]. Retroflexion also occurs across word boundaries, as illustrated in figure 2. [Spectrograms in this paper have an overlaid calibration grid with marks vertically at 1000 Hz intervals and horizontally at 100 ms intervals.] In figure 2, the word *or* which occurs in the phrase *votes or something* is embedded in frication. Note the continuation of frication throughout the [srs] interval and the very short interval of voicing. However, the neighboring [s]s both show substantially lowered resonant frequencies compared to this speaker's normal [s]. Like the contextual effects of vowel identity on the interpretation of stop formant transitions, retroflexion of [s] is a contextual effect. Though note that the example given in figure 2 poses a bit of a problem for Elman & McClelland's model, because there may be precious little evidence for the conditioning environment, the *or* in this phrase, separate from the contextually conditioned effect. That is, primary and secondary cues may change places, so that primary cues are no longer dominant.

Despite the potential difficulty with primary and secondary cues, contextual variability is a very tame example of variability in spoken language. In this paper I will suggest that "meaningful" variability poses a larger, more far-reaching, and hence more interesting problem for the theory of speech perception. Which is to say, I accept Elman & McClelland's interest in lawful variability and hope to push their strategy forward by examining other sources of variability. My approach is not compatible with an alternative perspective which seeks to ignore variability. This alternative is represented in part by researchers, for example Ken Stevens, who would argue that though lawful variability exists, there is no need to adopt complicated processing strategies like Elman & McClelland's because linguistic phonological features are characterized by invariant acoustic cues.

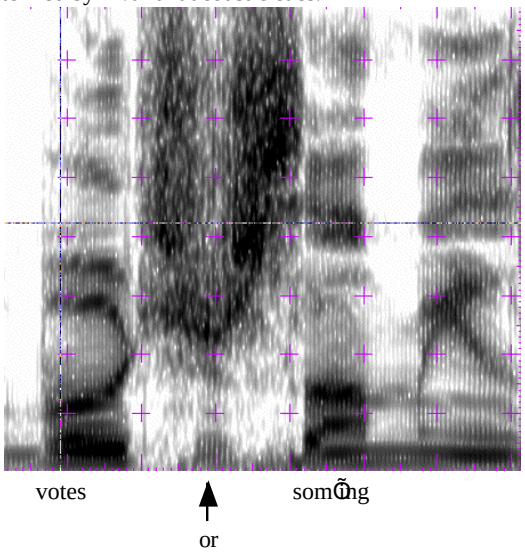

Figure 2. Contextual retroflexion in the phrase "votes or something".

Stevens' research project is a very important and powerful one, which seeks to find ways to identify significant acoustic properties that correspond to phonological natural classes. The foundational importance of this project for linguistic phonetics and phonology should not be underestimated. Nonetheless, I hope to make it clear why I do not share this perspective and why I believe that much valuable and interesting work opens up for us when we embrace meaningful variation as a research focus.

4. REDUCTION IN LINGUISTIC CONTEXT

The lawful variability addressed by Elman & McClelland's model is sometimes described as "mechanical", suggesting that coarticulation is caused by physical properties of the articulators. In contrast to this, Sapir ([8]) insisted that phonetic variation must be considered in linguistic context. By this he meant that a language's system of distinctive phonological contrasts delimits the range of acceptable phonetic variability to be found in that language. Thus, for Sapir patterns of phonetic variability are language specific.

We see evidence of this in a comparison of "perturbations of vowel articulations by consonantal context" in English and Korean (see figure 3). The effect of consonant context on [u] is much weaker in Korean than it is in English, as illustrated by the much smaller ellipse encompassing the Korean data. Note that [u] in coronal context in English is very similar to the Korean high central vowel [ö]. Sapir's account of the constrained, linguistically "meaningful", variability of Korean [u] is that the system of contrasts requires that [tu] not be confused with [tö]. There is no such constraint on American English [tu].

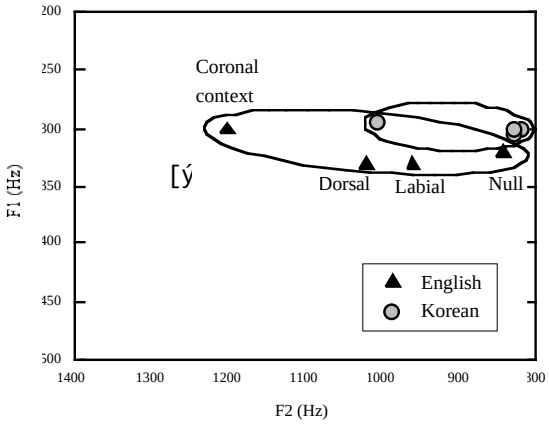

Figure 3. Consonant context effects on [u] in English and Korean. Vowel formant measurements from the vowel midpoint in [u] as a function of language (Korean versus American English), and preceding consonant context. The American English data come from Stevens & House ([8]), and the Korean data are from Jun ([5]).

5. FUNCTION WORDS

Beyond the system of contrasts that the lexicon defines, the larger syntactic and semantic context delimits a range of permissible speech reduction patterns. This is especially apparent in the realization of function words because they follow different patterns of reduction than do content words (Lee, [7], p. 100).

For example, the word *to* can be reduced to a stop release burst passing directly to a following [s] in, the phrase *one year I forgot to send it in*. In such a production the only acoustic evidence for *to* (which at a lexical level of reception is unquestionably present to native speakers) is a release burst and a bit longer [s] in the following word *send*. This is a typical example of a reduction.

Figure 4, on the other hand, shows an example of what I would call a reformulation. Here the word *to* is realized with an affricate [ts] preceding a vanishingly short schwa though there is no

/s/ in the following context. This token is from a talker who speaks very carefully and does not produce this affricated /t/ in other words. This seems like a reformulation or alternate coding of the function word, rather than a contextually driven coarticulation process.

Another phenomenon that we have observed with function words in the ViC database is called "gradient reduction". With reformulation it would be reasonable to propose that the word "to" simply has two stored forms /tu/ and /ts«/. However, an abstract representational solution cannot explain cases of gradient reduction. Consider for example the realization of *have* in figure 5. The utterance is *I have a very general...*. When listening to the whole context, native speakers report that the utterance starts with *I have*. However, if you isolate the first two words and play them out of context they sound very much like *I've*. Looking closely at the spectrogram we can see a slight dip in the amplitude of F3 which corresponds to a momentary change in phonation type;

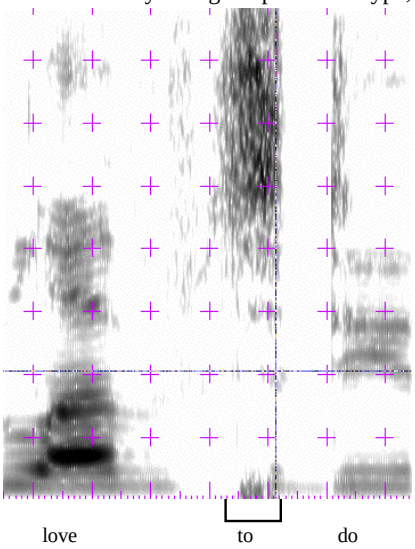

Figure 4. The word *to* is realized as [ts].

a momentary instance of breathy voicing. This slight change of phonation is a nod in the direction of /h/. The /h/ is not fully deleted, but it is also not fully realized - hence the ambiguity of the speech signal for native listeners. I would assert that what makes this sound like *have* in context is that the longer stretch of speech provides a rhythmic context in which it is apparent that a word "beat" occurred in that interval.

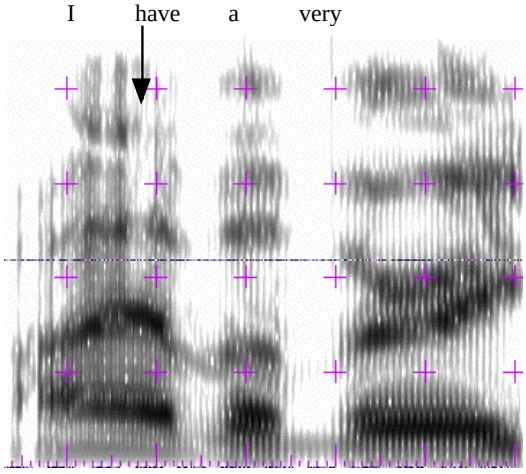

Figure 5. Cues, including F3 amplitude and rhythm, for the /h/ of "have" are weak, but present.

In a sense the rhythm of the utterance in figure 5 serves to cue the presence of the function word *have*. We can see a more extreme example of such a "rhythmic word" in figure 6. Here the word *of* occurs phenomenologically in the utterance *you know, and that was sort of another thing*. However, positive acoustic evidence for the existence of this word is mainly absent from the speech signal. Upon close inspection we find that there is a brief vowel (two glottal pulses) and a very

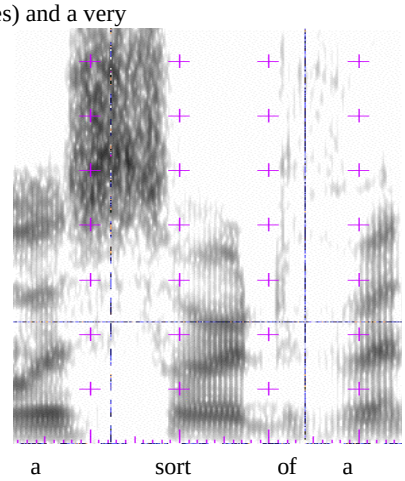

Figure 6. A "rhythmic word" in *sort of*.

low amplitude broad band fricative. But these elements of the signal are all but invisible in a spectrogram and inaudible in normal listening conditions. In fact you can hardly notice the difference when you replace the *of* portion of the signal with silence.

The examples discussed in this section illustrate that function words can be phonetically reformulated, may be gradiently reduced, and in some cases may be signaled primarily by rhythm. These observations can be taken as illustrative of the fact that variability in spoken language is guided by linguistic context. As in the case of variability guided by the system of contrasts in the phonological system, we see also that variability can be shaped by syntactic and semantic predictability (see also Jurafsky et al., [6]). These patterns of function word reductions are "meaningful" in that grammatical class is signaled by a pattern of variability unique to function words.

6. EMPHATIC PRONUNCIATIONS

In the ViC database there are numerous instances of highly emphatic productions in which the talker exaggerates some aspect of a word pronunciation.

Figure 7 illustrates this with a spectrogram of a production of the word *repealing* in which the [p] closure is over 150 ms long, which is 2 to 3 times longer than the other closure intervals in this stretch of speech. Although, the phonetic information in this production of *repealing* is roughly consistent with that found in less marked productions of the word, the abnormally long closure interval is striking, and could be enough to signal a boundary to listeners or recognition algorithms. Why would a talker complicate lexical access for the listener in this way? The long [p] closure interval in this token serves (along with heavy aspiration at the release) to convey emphasis. This instance of increased variability serves to convey nonlexical meaning, and help the listener construct an accurate representation of the prosodic structure of the utterance. In addition to contributing to the conveyance of prosodic structure, emphatic pronunciations add a level of gradient salience to words. Together with prosody this directly contributes to the listener's on-line construction of pragmatic discourse structure.

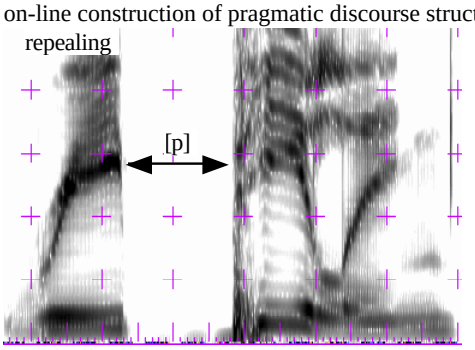

Figure 7. An emphatic production of *repealing* with an unusually long [p] closure interval.

It is sometimes said that prosody provides a parallel channel of communication in which the talker "comments" on the lexical message. For example, talkers can use intonational tunes to make statements with a stronger or weaker degree of finality. Emphatic pronunciations seem to serve a similar function, so that a word is given an additional "meaning" such as "this is the main point of what I'm trying to say", or "this point contrasts as an opposite to another point recently made in the discourse". This is important because I am suggesting, as Mary Beckman has suggested many times previously, that spoken language is prosody all the way down. If this is so, then segmental variability is prosodically meaningful; not merely random noise, or mechanically "lawful" coarticulation.

7. VARIANT PRONUNCIATIONS THAT WORK

Figure 8 shows a spectrogram of the word "extraterrestrial". On first hearing it is apparent that the talker's pronunciation of this word is quite a bit different from the dictionary pronunciation, however most of the difference seems to come from the end of the word which is pronounced something like [trES«l]. However, the first cluster of consonants also shows a unique pronunciation. Instead of the canonical sequence /kstr/ this pronunciation has the sequence [StSr]. We aren't sure whether this pronunciation is typical of a community of talkers in central Ohio. It could be. One way to characterize a unique pronunciation such as this, is as a "groping" pronunciation. The talker has heard the word, but it is a long, uncommon word, so the talker "gropes" for the correct pronunciation.

One of the most interesting things about groping pronunciations is that listeners usually do access the correct word from the lexicon even though the pronunciation may be articulatorily quite off target. For example, in *extraterrestrial* the velar stop closure of /ks/ has been eliminated, and the /s/ is realized as [S]. Note that [S] is spectrally more like a [k] release than is [s]. How then do listeners access the correct lexical item for this pronunciation? Groping pronunciations may pose a problem for lexical access models that assume a strict separation of processing levels so that phoneme perception is completed prior to lexical access. One strategy has been to use underspecified phoneme representations (for example, mid-class phonetic

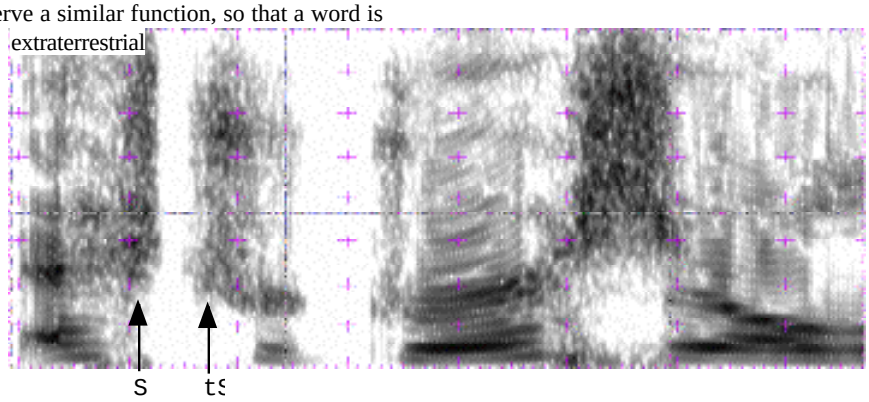

Figure 8. A pronunciation of *extraterrestrial* that seems to work. Note that the /kstr/ of the canonical form is [StSr].

categories, Dalby et al., [1]) as input to lexical access. However, in *extraterrestrial*, [S] and [ks] don't match even at this level. Clearly such models must allow some phonetically nuanced information to feed up to the lexical access procedure.

In keeping with the theme of "meaningful" variation, I would also point out that alternate pronunciations of words form the basis for the identification of dialect variation. So, that though variant "groping" pronunciations are to an extent arbitrary, they can be meaningful to listeners as markers of personal identity or group membership.

8 IDIOSYNCRATIC PHONETIC SPACE

Another kind of variability that we have found in the ViC corpus is fairly straightforwardly idiosyncratic. For example, figure 9 shows an instance of the phrase *township trustee*. This spectrogram shows an [s] in *trustee* that has substantial spectral energy extending down to the frequency range for F3, like the fricative noise for [S] in *township*. Despite their spectral and auditory similarities they remain noticeably different with [S] showing slightly lower frication noise and less energy above 6 kHz. The spectral cues distinguishing [s] and [S] in other talkers' productions are consistent with these differences, [s] has a higher spectral center of gravity. What is idiosyncratic for this speaker is that the spectrum for [s] has more low frequency noise than usual.

Figure 9 is reminiscent of Sapir's ([7]) discussion of "the intuitive 'placing' of the sounds with reference to one another". In this discussion he touches on individual variation.

> "It is true that no two individuals have precisely the same pronunciation of a language, but it is equally true that they aim to make the same sound discriminations, so that if the qualitative differences of the sounds that make up A's pattern from those that make up B's are perceptible to a minute analysis, the relations that obtain between the elements in the two patterns are the same."

The relational invariance that Sapir appeals to is directly analogous to ratio coding schemes of vowel normalization that have been proposed in more recent decades. But the case of *township trustee* does not quite correspond to a ratio coding scheme, because it is not as if the playback speed of the recording has been altered (the typical situation inviting of a ratio scheme). Instead the talker simply has a somewhat idiosyncratic pronunciation of only one phoneme [s]. To "normalize" to this talker's production of [s] the listener needs to warp only a small portion of the perceptual space.

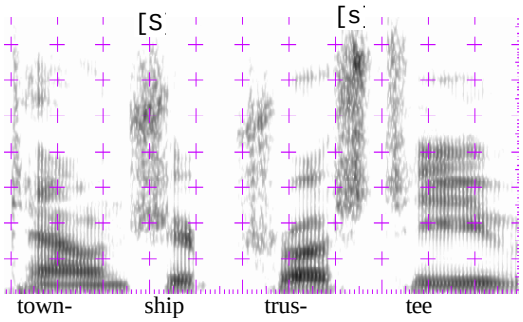

Figure 9. A production of the phrase *township trustee* showing that [s] is spectrally similar to [S] for this talker.

9. MEANINGFUL VARIATION

The theme of this paper has been that variability in spoken language is "meaningful". I have shown examples from unmonitored conversational speech that illustrate this point at a number of levels of analysis - that variability is shaped by a language's system of contrasts, the syntactic function of words, or the emphasis that a talker wants to convey. The examples in the last two sections also highlighted the fact that individuals differ. Johnson, Strand & D'Imperio ([4]) put forward the idea that people "perform" gender in their speech. In keeping with this perspective I would like to conclude with an example of spoken language variability that arises from this performance aspect of ordinary speaking.

The phrases shown in figure 10 were spoken in the context of a young mother talking about how people dress their children. She said, "...put a sweater on her on Tuesday, and a little skirt on her on Wednesday." The words by themselves don't convey the talker's meaning in this utterance. The first phrase is spoken rather softly

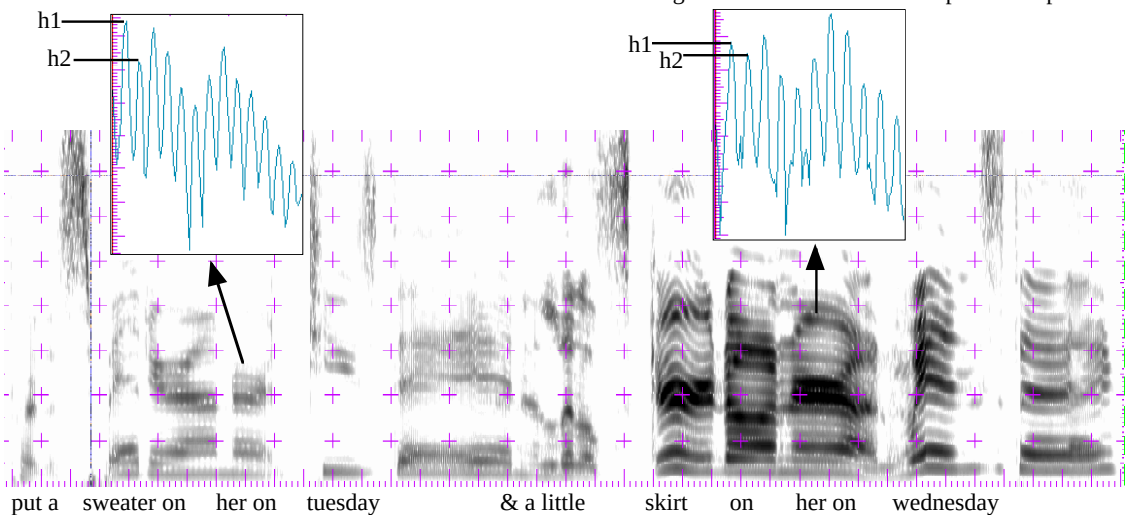

Figure 10. Phonation type, F0 range, and loudness marking contrasting phrases in conversational speech. The spectra were calculated at the mid point of the vowel in *her*.

and with a breathy phonation type (as indicated by the ratio of the first two harmonics). The second phrase is spoken more loudly, with a higher F0 and with a more creaky phonation. The effect of these changes is to indicate that the talker is making fun of the behavior she is describing - treating a child like a dress-up doll. This information is conveyed by the phonetic differences between the two phrases in this utterance.

People use subtle phonetic cues to convey nuanced information about the communicative context when they talk to each other. Thus, in contrast to the traditional view that speech perception theory merely needs to account for phoneme recovery, this look at conversational speech suggests that variability in spoken language is not merely random noise as a simple linear associator model would assume, and it is not merely "lawful" as a mechanistic phoneme transmission view would assume. Rather,

spoken language varies as talkers blend together complex linguistic and supralinguistic messages. This rich "meaningfulness" of variability is a worthy focus of study which presents some interesting deep challenges to conventional theories of speech perception and auditory word recognition.

ACKNOWLEDGEMENTS

The Variation in Conversation corpus is the result of a joint collaborative project of the Departments of Linguistics and Psychology at Ohio State University. Funding was provided by Grant No. R01 DC04330-01 from the National Institute of Deafness and Ohter Communication Disorders, to Mark Pitt, Elizabeth Hume, and Keith Johnson. Personnel on the project include Scott Keisling, William Raymond, Jennifer Muller, Matt Makashay, and Robin Dautricourt.

Keith Johnson, 222 Oxley Hall, 1712 Neil Ave., Columbus, OH 43210. Email: kjohnson@ling.osu.edu.

REFERENCES

[1] Dalby, J., Laver, J. & Hiller, S.M. (1986) Mid-class phonetic analysis for a continuous speech recognition system. *Proceedings of the Institute of Acoustics*, 8, 347-354.

[2] Elman, J. & McClelland, J. (1986) Exploiting lawful variability in the speech wave. In Perkell, J.S. & Klatt, D.H. (eds.) *Invariance and variability in speech processes* (pp. 360-380). Hillsdale, NJ: Lawrence Erlbaum Associates.

[3] Goldstein, L. (1983) Vowel shifts and articulatory-acoustic relations. *10th International Congress of Phonetic Sciences*, pp. 267-273.

[4] Johnson, K., Strand, E.A., D'Imperio, M. (1999) Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*. 27, 359-384.

[5] Jun, S-A. (1994) Labial position and acoustics of Korean and English high vowels. *OSU Working Papers in Linguistics* 43, 70-84.

[6] Jurafsky, D., Bell, A., Gregory, M.L. and Raymond, W.D. (2000) Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In Bybee, J. & Hopper, P. (eds.). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

[7] Lee, Kai-Fu (1989) *Automatic speech recognition: the development of the SPHINX system*. Norwell, MA: Kluwer Academic.

[8] Sapir, E. (1925) Sound Patterns in Langauge. *Language* 1, 37-51.

[9] Stevens, K.N. & House, A. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hhearing Research* 6, 111-128.