TR-IT-0115

Laryngeal Correlates of Local and Global Vocal Prominence

Mary E. Beckman

May, 1995

ABSTRACT This paper describes three speech databases that were collected in collaboration with colleagues in ATR-ITL and ATR-HIP. Each database includes one or more other signals recorded synchronously with the audio – ranging from non-invasive laryngograph trace to muscle activity level measured from subcutaneous electrodes. The corpora were designed to explore laryngeal correlates of prominence relationships in English at several levels, from relative overall loudness of the utterance as a whole to relative stress of a syllable within a word. In between these two sizes of unit, are prominence relationships for different words within a phrase or for different phrases within the utterance. Studying these intermediate levels is complicated because, in most languages, increased prominence can raise the average voice fundamental frequency over the word or phrase, but raised pitch is also a reliable effect of increased vocal effort over a whole utterance. In languages such as English, studying these phenomena of word-level or phrase-level prominence relationships is further complicated because of the way that they interact with the syllable-level prominence relationships, which are less directly related to pitch raising than in Japanese. The three databases were designed to address questions at each of these levels. For each database, I will outline the background issue and (where available) describe relevant analyses.

1. Introduction.

In this paper, I will describe databases of recordings of three speech corpora that were collected in collaboration with colleagues in ATR-ITL and ATR-HIP. The corpora were designed to examine some of the phenomena that are important for our understanding of prominence relationships in English utterances.

Getting a full understanding of prominence relationships involves questions at many levels, along several different dimensions of analysis. One of the most important of such dimensions is the size of the speech units involved in the prominence relationship. At one end of the scale, we can compare the relative prominences of different speaker turns. In a conversation between two people, one conversational participant may speak with a louder voice than the other, for any number of reasons. The speaker may be overcompensating for a slight deafness, or may be trying to project above noise in the environment, or may be emotionally involved in the topic of the conversation — shouting because of anger or excited happiness. The softer voice of the other turns, conversely, may be because the speaker is doctoring a sore throat, or is concerned to not let a third party overhear the conversation, or is conveying emotional detachment in an attempt to calm the louder speaker. We addressed questions at this level in the productions of utterances in the first and third databases, by asking the speaker to vary the overall prominence of her voice across different utterances of the same sentence type, by acting out different degrees of "speaking up" to project above environmental noise.

At the other end of the scale, we can compare different syllables. In languages such as English, in particular, this is a linguistically important level, because prominence at this level is part of the lexical specification of polysyllabic words. For example, the bisyllabic words *insight* and *incite* contrast minimally in whether the first or the second syllable is more prominent. Similarly, the four-syllable words *legislature* and *legislation* contrast in whether the first or the third syllable is more prominent, although in both cases these two syllables are both more prominent than the second and fourth syllables. The phonetic correlates of the prominence relationships among syllables are notoriously difficult, and we will return to them below. We tried to address some of these questions in designing the second and third corpora, by including target words that vary in which syllable is stressed — i.e. which syllable is lexically specified as most prominent.

In between these two sizes of unit, we can talk about the relative prominence of different words within a phrase or of different phrases and sentences within a speaker's turn. Understanding prominence relationships at these levels is very important for all languages, since here prominence is closely related to such phenomena as discourse topic structure, old versus new information, and focus of attention. For example, one word might be more prominent than another in its phrase because the speaker wants to draw narrow focus of attention onto that word, or one phrase might be more prominent than another in the speaker's turn to signal that it introduces a new topic or subtopic into the discourse flow. We tried to address questions at these levels in all three of the corpora, by constructing a dialogue context for each utterance type (or by having the speaker imagine a dialogue context, in the recordings of the third corpus) that varied the discourse role of different phrases or put narrow focus on one or another word.

Studying prominence relationships at these intermediate levels is very complicated because one of the most reliable acoustic effects of increased prominence is a raising of the average voice fundamental frequency over the whole unit. This effect of "expanded backdrop pitch range" is common to all of these levels and also to the overall level of vocal effort. That is, backdrop pitch range can be increased over a word when the speaker wants to put narrow focus of attention

on that word. (This is particularly true of Japanese — see Pierrehumbert & Beckman, 1988; Maekawa, 1991, 1995; Tsumaki, 1994; Fujisaki, Ohno, Osame, Sakata, & Hirose., 1994). Similarly, backdrop pitch range can be increased over a larger phrase when a speaker wants to signal that the phrase introduces a new topic of conversation. (This seems to be true of many languages — see, e.g., Silverman, 1977, for English; Avesani & Vayra, 1988, for Italian; Swerts & Geluykins, 1994. for Dutch; Nakajima & Tsukada, 1995, for Japanese.) But backdrop pitch range will also increase over the whole turn when the speaker talks more loudly to project over background noise (e.g., Liberman & Pierrehumbert, 1984) or in a loud angry voice (e.g., Higuchi, Hirai, & Sagisaka, 1994). It is difficult to separate these different domains of pitch range expansion in an acoustic analysis of fundamental frequency, because the analysis is highly dependent on the control model for the backdrop pitch range at each of the linguistic levels assumed, yet it is impossible to know a priori what is the best model at each level, or even how many levels of prominence relationship should be linguistically distinguished by identifying different independent control units. For example, Fujisaki's model for pitch range relationships in Japanese (Fujisaki & Sudo, 1971) posits an "accent command" (to model the backdrop pitch range of an accentual phrase) and a "phrase command" (for the backdrop pitch range of a higher-level unit of grouping). Hirai, Higuchi, & Sagisaka (1994) have successfully used an analysis-by-synthesis technique to extract these two control parameters semi-automatically in order to study the effects of different emotional contexts.[1] Takeda and Ichikawa (1990) used a similar technique to examine the effects of putting narrow focus on word in the utterance. However, it is not simple to generalize the model and the technique to examine analogous prominence relationships in English, because English has no linguistic unit comparable to the accentual phrase in Japanese (see, e.g., Beckman & Pierrehumbert, 1986; Venditti, Jun, & Beckman, in press).

In languages such as English, studying these phenomena of word-level or phrase-level prominence relationships is further complicated because of the way that they interact with the syllable-level prominence relationships. For example, when one word is made more prominent than another in the phrase, that prominence affects different syllables within the word differently. The stressed syllable (i.e., the syllable that is lexically specified as the most prominent within the word) might be substantially higher than any following syllable. Moreover, the pitch difference can be larger or smaller, depending on how much more prominent the word is than its neighbors. This is reminiscent of the pitch range expansion for a phrase as a whole when it is made more prominent than neighboring phrases. However, the two phenomena differ in two ways. First, when a word is relatively prominent, the pitch expansion does not cover the whole word but is localized to the lexically prominent syllable. Also, the lexically prominent syllable can be substantially lower in pitch rather than higher in pitch than surrounding syllables, depending on the pragmatic context. (See Chapter 1 of Ladd, 1980, or Beckman & Ayers, 1984, for tutorial descriptions of these facts.)

These two differences between phrasal prominence and word-level prominence in English occur because words that are prominent in a phrase typically will be produced with a pitch accent, and the pitch accent will be phonologically associated to the lexically prominent syllable . The pitch accents of English are not like the pitch accent of Japanese. Rather than there being a single falling type (as in lexically accented words in Tokyo Japanese), there are six different pitch accent types, which contrast tonally in whether the accented syllable is high or low and whether the pitch around it is level, rising, or falling. The accents signal complex

---

[1]The study also extracted "$F0_{min}$" to show that speaking in an angry voice affects this more global pitch range control as well.

pragmatic relationships between the word and the background information in the discourse (see Pierrehumbert & Hirschberg, 1990). Thus when the pragmatic context calls for a high pitch accent on the focused word the lexically prominent syllable typically will be much higher in pitch than surrounding syllables, whereas when the pragmatic context calls for a low pitch accent, the lexically prominent syllable typically will be much lower in pitch than following syllables. Since the relatively greater prominence of accented syllables can involve either relatively higher pitch or relatively lower pitch, it is clear that greater prominence at the word level cannot be simply a matter of expanding backdrop pitch range by boosting the accent command. Thus, English is very different from Japanese, where the relative prominence of an accentual phrase is signalled primarily by its overall backdrop pitch range (see, e.g., Maekawa, 1995).

The difference between English and Japanese accent is even more striking when we consider native speaker intuitions. The intuitions of native speakers of Japanese match the known phonetic correlates of lexical accent in the language. When Japanese-speaking phoneticians describe accent patterns, they almost invariably describe them in terms of the pitch pattern. By contrast, native speakers of English have strong intuitions that lexical "stress" involves loudness. For example, Henry Sweet defined it as follows:

> Physically force is synonymous with the effort by which breath is expelled from the lungs.... Acoustically it produces the effect known as 'loudness' which is dependent on the size of the vibration-waves which produce the sensation of sound.... The comparative force with which the syllables that make up a longer group are uttered is called 'stress'. [Sweet, 1906, pp. 47 & 49]

That is, native speaker intuition identifies increased loudness and not increased pitch as the psychoacoustic correlate of lexical accent in the language.

However, it has been extremely difficult to document any phonetic basis for this intuition other than the negative fact that accented syllables are sometimes lower in pitch rather than higher in pitch. Classic experiments on the production and perception of lexical stress have shown overall intensity to be a very weak and extremely unreliable cue to differentiating stress pairs such as *insight* versus *incite*. For example, Fry (1955; 1958) showed that overall RMS amplitude is not an effective heuristic for distinguish such words, and that boosting the overall RMS amplitude of the first or second vowel in such word pairs is not a very effective way to shift the perceived stress pattern. It is much more effective to lengthen the vowel, and it can be even more effective to manipulate the fundamental frequency pattern so as to change the perceived pitch accent placement, although the effectiveness of this manipulation will depend upon how well the result mimics the target intonation pattern (see Beckman, 1986, chapter 3, for a discussion of this and other related experiments).

Beckman (1986) proposed that these classic experiments failed to uncover reliable acoustic measures corresponding to native speaker intuition because of a misunderstanding of the psychoacoustics of loudness. That is, psychoacoustic experiments have shown that the relationship between signal intensity and perceived loudness is much more complex than the relationship between signal frequency and perceived pitch. The loudness of a signal with a given overall RMS intensity can be greatly affected by other aspects the signal. For example, a complex harmonic signal (such as a vowel) can be more or less loud, depending on the distribution of the intensity in the frequency domain. This dependency is encoded in standard loudness measures such as ISO532 (1975). Another effect that is not encoded in any standard loudness measure is that, for signals shorter than about 200 ms, loudness depends on signal length. Other things being equal, a

longer signal is louder. Beckman (1986) and Beckman & Edwards (1992) proposed that this effect of "temporal summation of loudness" is the basis for the documented effects of stress on syllable duration, and the perception that longer syllables are stressed.

Sluijter similarly has proposed that stressed syllables will be measurably louder if we only look at the right measure of loudness (Sluijter & van Heuven, 1993; Sluijter, 1994). Her work concentrates on the frequency effects. She found differences in spectral tilt corresponding to syllable-level prominence in Dutch (a language which has a stress-accent system essentially like that of English). Accented syllables reliably have more energy in high-frequency regions, where the dynamic range for loudness perception is largest. She relates this finding to Gauffin & Sundberg's (1989) results showing that when speakers use increased vocal effort over the entire utterance to project a louder voice, there is an increase in the speed of glottal closing, and a consequent increase in energy that affects frequency bands above 1000 Hz much more than it affects the bands just above the fundamental frequency. Campbell (in press) points out that the existence of these spectral correlates of prominence has profound consequences for the design of concatenative speech synthesis systems if we want to achieve natural sounding prominence manipulations across the hierarchy of levels of prominence control.

To summarize this background motivation, then, we want to understand prominence relationships among linguistic units at different levels in English. To do this we clearly need to study fundamental frequency patterns, to understand where differences in backdrop pitch range come from and how they interact with the complexities of pitch accent in this complicated intonational system. However, F0 clearly is not the only aspect of the signal that is important. Even if we look only at aspects of the signal that are related to laryngeal control, there are many questions we need to address concerning the spectral correlates of increased vocal effort that Sluijter and van Heuven describe. The three databases described below were designed to address some of these issues.

The databases vary in the range of phenomena examined and in the types of signal that were recorded for the different utterance types. Recordings for the second corpus include utterances by three speakers, but the utterance types were produced in only one overall vocal effort level, and we recorded only the audio signal and (for about a third of the utterances) an accompanying laryngograph signal. The first and third databases have utterances by a single speaker, but produced in several overall prominence levels. Also, in addition to the audio recording for these corpora, we have accompanying synchronous recordings of electromyographic activity level measured from electrodes inserted subcutaneously into several intrinsic and extrinsic laryngeal muscles, and also (for a third of the utterances in the third database) synchronous recordings of subglottal pressure. We began recording these other physiological signals in earlier databases, because we thought that looking at them might give us an insight into the paradox that accented syllables can be lower in pitch than surrounding syllables and still be perceived as more stressed. We also found, however, that the physiological signals can help us disentangle the different levels of prominence relationship for high pitch accents. In each of the following sections, therefore, I will begin by describing the specific background questions motivating the particular corpus and the choice of signals to record.

2. The "lean mini-noodle" corpus

2.1. Background. This corpus was used throughout a series of experiments done in collaboration with Kiyoshi Honda of ATR-HIP (now at the Waisman Center at the University of Wisconsin) and Donna Erickson (while she was a

visiting researcher at ATR-HIP). Other people who have been involved in this collaboration are Hiroyuki Hirai of ATR-HIP (now back at the Sanyo Hyper-Media Research Laboratory) and Seiji Niimi (director of the Research Institute of Logopedics and Phoniatrics, University of Tokyo). We have recorded and analyzed repeated utterances of the sentences listed in Figure 1, produced with the intonation types described in the figure, in each of three different speaker-determined levels of overall vocal effort, ranging from a softer than normal voice to a louder than normal voice.

1. Nuclear L* pitch accent, in the canonical yes-no question contour:

```
What would you like for lunch?
D'you have a lean mini-noodle dish?
                  L*          L*    H- H%
```

2. Nuclear L*+H scooped pitch accent, in a contour indicating pragmatic uncertainty:

```
Do all of your rice dishes have this fatty meat sauce?
We have a lean mini-noodle dish.
             H*          L*+H   L- H%
```

3. Nuclear L+H* rising peak accent, in the canonical contrastive emphasis contour:

```
Do you have any bean dishes other than this couscous thing?
We have a lean mini-noodle with beans.
             H*      L+H*     L-        L%
```

4. Sentence-medial L- phrase accent marking the boundary between two intermediate phrases:

```
Do you have any pasta less fattening than fettuccine
  Alfredo?
We have a lean, mini-noodle dish.
             H* L-        H*    L- L%
```

5. Pre-nuclear H+L* pitch accent, in a contour indicating an obvious pragmatic inference (here of resignation):

```
Edward, you know we're not supposed to eat meat at
  lunchtime.
Oh, all right. We'll have the lean mini-noodle with beans.
                       L*        H*         H+L*       H+L* L- L%
```

Fig. 1. Discourse contexts and intended intonation contours for the five sentence types. See Appendix A for a more detailed description.

The original motivation for designing this corpus of sentences was to examine the behavior of various low tones occurring in different pitch accents or at the edge of an intermediate-level prosodic phrase in English. Current models of intonation with any reasonable coverage of prominence-related phenomena are based on extensive examination of the behavior of F0 peaks. The relevant research shows that fundamental frequency values of H tones (local targets high in the pitch range) can be predicted by assuming a fairly simple interaction between the values inherent to different local tonal commands (associated with peak accents and H edge tones) and the variable specifications of more global backdrop pitch range values that can be associated with degree of phrasal prominence or overall vocal effort. For example, Liberman and Pierrehumbert (1984) showed that, in a common English "downstepping" contour, each successive H-toned accent target is proportionately lowered relative to the preceding pitch accent, and this proportion is constant over different global pitch ranges for soft, normal, and loud voice. By contrast, much less is known about L tones (targets low in the pitch range). In English intonation, these include the L* nuclear pitch accent on the most stressed syllable in the "yes-

no question contour" and the L- phrasal tone in the "declarative contour" that simultaneously marks the end of the phrase and the fall from the peak accent.

The particular intonation contours in Figure 1, therefore, were chosen to exemplify as many different types of low tone as possible. In the transcriptions of the five "tunes" in the figure, these target L tones are underlined.[2] The segmental "text" of the sentence for each tune was chosen so as to put the target L tone in a fixed context that would simultaneously minimize segmental effects on the F0 contour and also on sternohyoid activation level, which we recorded in all but two of the experiments using this corpus. The strap muscles, and the sternohyoid in particular, are fibers running vertically just underneath the surface at the front of the neck, which have been shown to be recruited in producing very low F0 values in many languages, apparently because of their role in lowering the larynx. (See Honda, Hirai, & Kusakawa, 1993, for a plausible recent explanation of the mechanism.) Thus, we wanted to be able to examine F0 minima associated with the target L tones without worrying about the well-known effects of obstruent consonants on fundamental frequency (e.g., Lehiste, 1970; Silverman, 1987), and we wanted to examine sternohyoid activity associated with lowering the larynx to produce low pitch without worrying about the interaction with the use of this muscle in opening the jaw to produce low vowels (see Sawashima, Hirose, Yoshioka, Horiguchi, & Kiritani, 1983; Yoshida, Honda, & Kakita, 1992). Results from earlier recordings of this corpus are reported in Erickson, Honda, Hirai, & Beckman (1993; 1995) and Erickson, Honda, Hirai, Beckman, & Niimi (1994), the second of which is attached as an appendix to this technical report. In those papers, we describe analyses of the target L tones in utterances by three speakers, for whom we recorded sternohyoid muscle activity or (for one set of recordings) subglottal pressure.

The database that I will describe here is the latest in this series of experiments. The recording was made in December 1993 at the RILP in the University of Tokyo Medical School when I was visiting Japan for ten days to attend a symposium at Dokkyo University, and the speaker was me, a female native speaker of American English.[3] In this latest experiment, we recorded EMG muscle activity from the sternohyoid muscle (SH), the anterior belly of the digastric muscle (ABD), and the cricothyroid muscle (CT), using hooked-wire electrodes inserted subcutaneously into the neck. The ABD we recorded because it, like the SH, is regularly used in opening the jaw, and we wanted to be able to sort out SH activity related to low vowel production from SH activity related to low tone production. The CT we recorded because it is the primary muscle involved in laryngeal adjustments to raise pitch. Although we recorded only from one side, the CT is actually a pair of muscles, whose fibers run vertically and obliquely on both sides of the larynx, from the lateral interior surfaces of the thyroid cartilage to the lateral superior surfaces of the cricoid cartilage. Contraction of these fibers can pull down the front

---

[2]Note that in the labelling system used here at ATR-ITL, the target tone in the last utterance type would be transcribed as a downstepped high tone. That is, H+L* in the figure corresponds to H+!H* in the ToBI system — see Beckman & Ayers, 1994.

[3]During that trip, we also recorded another set of my productions here at ATR-HIP so that we could have a subglottal pressure (P_0) signal for the same speaker in utterances of the same corpus produced at about the same time. In our most recent paper, attached as Appendix B, we then compared the pattern of cricothyroid muscle activity with the pattern in the P_0 signal during the production of the relevant H tones. (The P_0 pattern associated with the L tones in this database was described in our earlier ICSLP paper, attached as Appendix A.)

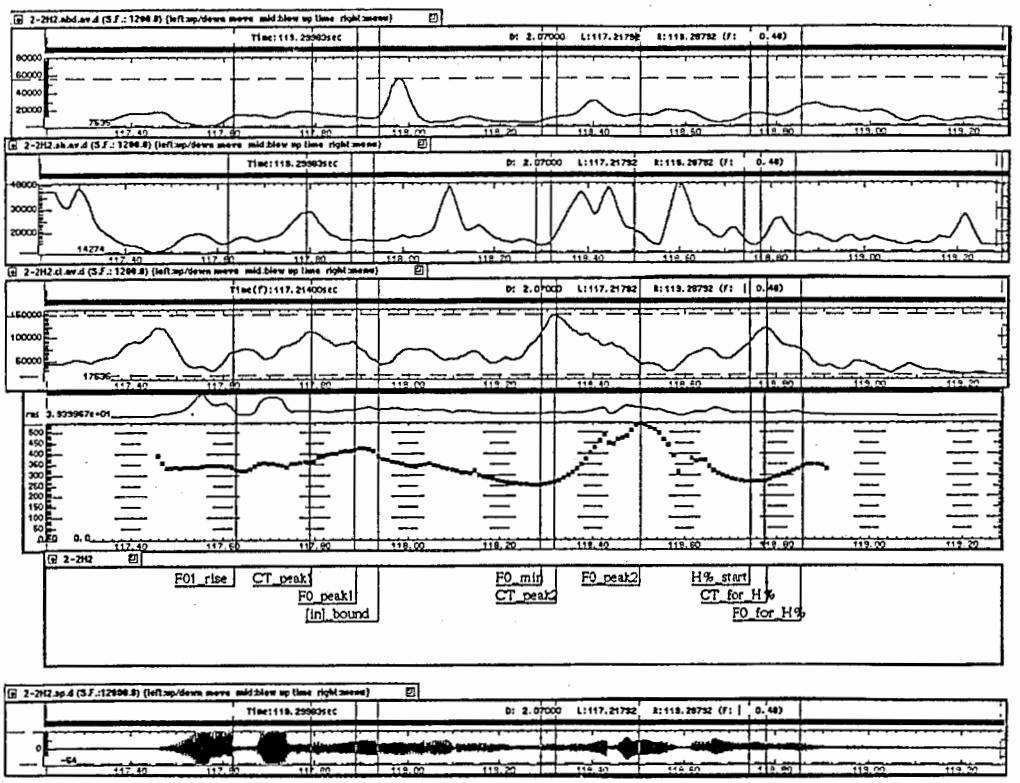
of the thyroid cartilage, elongating the vocal folds to raise pitch. Since the muscles are very small, and the electrode must be inserted through a very narrow opening between the two cartilages, it is not easy to get a good signal. However, Dr. Niimi managed to get a very clean signal after trying only two insertions.

The fact that we had such a good CT recording, together with the fact that the tunes for utterance types 2, 3, and 4 create peaks for different kinds of H targets around the target L tone, convinced us that this database would be useful for looking at the behavior of H tones as well as the originally targeted L tones. Donna Erickson was a visiting researcher in ATR-HIP for several months in 1995, and so I brought this set of recordings back with me to Japan, so that she, Kiyoshi Honda, and I could analyze the CT signal in October, 1995, when the last month of her visit overlapped with the first month of my stay at ATR-ITL. This set of utterances then constitutes the database that is now available online in directory `/usr/pi/data/RILP93`.

2.2. The database. The structure of database `/usr/pi/data/RILP93` is as follows. The data for each utterance are stored separately in files with a common basename. An example basename is `2-2H2`. The significance of the name is as follows. The first digit (before the hyphen) is the order number; the second digit (after the hyphen) is the repetition within the order; the capital letter is the `Low`, `Normal`, or `High` vocal effort level; and the last digit is the utterance type, from 1 to 5, as in Figure 1. Thus the basename `2-2H2` means order number 2, repetition number 2, with High effort (i.e. loud voice), for utterance type 2, and `4-8L1` means order 4, repetition 8, with Low effort (i.e. soft voice), for utterance type 1, and so on. A word is in order about the order numbers. I read the list of utterance types in three different orders, with eight repetitions at each order before going on to the next order, but they are numbered from 2 to 4 in the basenames, rather than from 1 to 3. This is because order 4 is actually order 1. After the first time through the second order, Dr. Niimi noticed that the SH signal was not very strong during low-pitched parts of the utterance. So he made a new insertion, and the signal before is not equivalent to the signal after. Therefore, we decided to discard the earlier readings in processing the data. However, at the end of the eight times through the third order, the signals from the electrodes were still good (meaning that blood had not yet started to clot at the tip of the electrode to impede the transmission of EMG activity level). So I went on to produce another set of 8 repetitions for the first order. Thus, we have up to 23 repetitions of each utterance type in each of the three vocal effort levels.

There are eight data files and a label file associated with each basename, distinguished by the following extensions:

`2-2H2.sp.d`	the audio file
`2-2H2.sp.f0`	the associated F0 file (generated with `get_f0`)
`2-2H2.sh.d`	raw SH EMG activity level
`2-2H2.sh.av.d`	smoothed SH
`2-2H2.abd.d`	raw ABD EMG activity level
`2-2H2.abd.av.d`	smoothed SH
`2-2H2.ct.d`	raw CT EMG activity level
`2-2H2.ct.av.d`	smoothed CT
`2-2H2.lab`	xlabel file



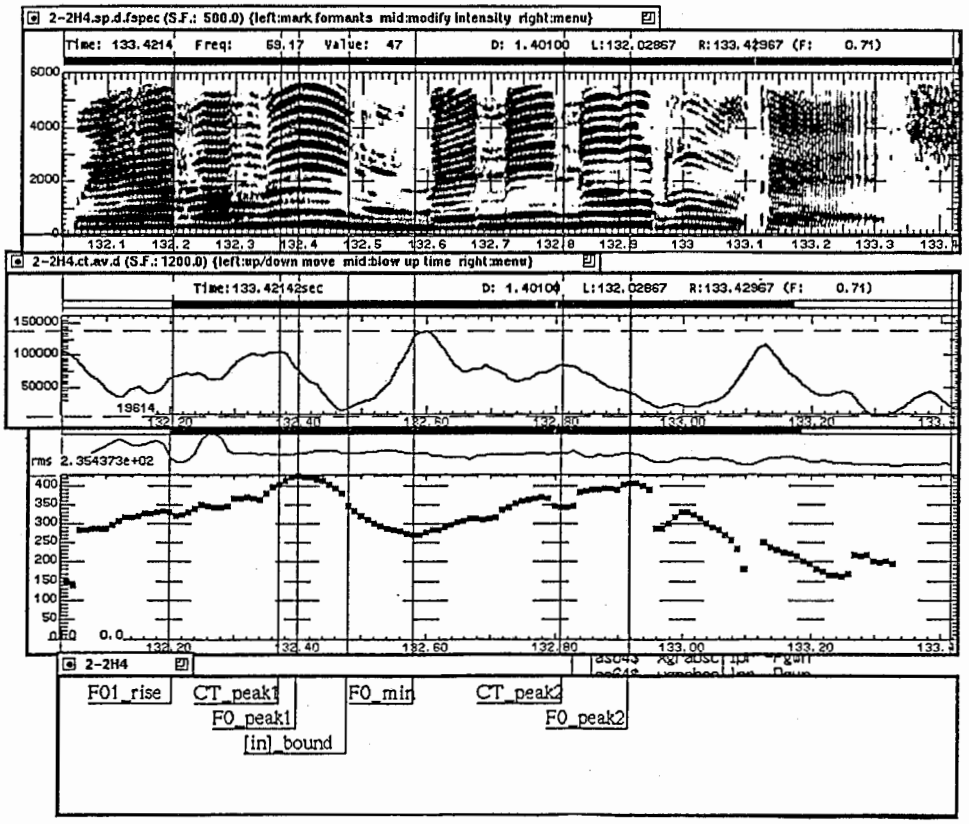


Fig. 2. Two sample displays from `show-emg`.

The smoothed data files such as `2-2H2.abd.d` contain the same EMG signal trace as the corresponding raw data files, except that the data have been downsampled to 1.2kHz from the overly high (for EMG) sampling rate of 12kHz, rectified, and passed through a moving triangular filter of 70 ms. All of the data files have ESPS headers, so that the data can be viewed in time-aligned windows using `waves+`. In the directory there is also a shell script called `show-emg` that displays the data relevant for analysis, in a layout of windows that I find easy to look at. The syntax for using the shell script is:

```
show-emg BASENAME1 BASENAME2 ....
```

The top part of Figure 2 shows a sample display for utterance `2-2H2`. In this utterance type, the fall after the first peak and subsequent rise to the second F0 peak is due the L+H* accent type. The bottom part of the figure shows another sample display where the user has then made the default `waves+` wide-band spectrogram to cover up the windows for the ABD and SH traces. The utterance here is of a different type, `2-2H4`, where the L is a phrase edge tone and second peak is a simple H* accent. A comparison of the F0 trace windows shows clearly the different relationships between the two F0 peak values in the two different utterance types. In both displays, the two F0 peaks for the target H tones are marked, with labels inserted at time positions chosen by Dr. Erickson. The other labels are for various other events that we also found useful to mark for our analyses. Figure 3 shows the label file for `2-2H2`, with my comments added to the right to describe the labeled events.

```
signal 2-2H2
type 0
color -1
comment created using xlabel Mon Aug 15 16:33:01 1994
font -*-times-medium-r-*-*-17-*-*-*-*-*-*-*
separator ;
nfields 1
#
  117.623917    -1 F01_rise     ; beginning of rise to first F0 peak
  117.792883    -1 CT_peak1     ; CT peak just before first F0 peak
  117.891640    -1 F0_peak1     ; first F0 peak (before the target L)
  117.938250    -1 [in]_bound   ; segment boundary before [n] of "lean"
  118.290250    -1 F0_min       ; F0 minimum for target L tone
  118.322220    -1 CT_peak2     ; CT peak just before second F0 peak
  118.505708    -1 F0_peak2     ; second F0 peak (after the target L)
  118.746675    -1 H%_start     ; beginning of rise to H% (for type 2)
  118.782428    -1 CT_for_H%    ; CT peak just before H% at end
  118.853533    -1 F0_for_H%    ; F0 peak at end of types 1, 2, 3
```

Fig. 3. Sample label file with comments added.

The directory also includes a subdirectory called PROGRAMS where I have left various shell scripts that we wrote to extract the relevant data values in batch using the label files. For example, the shell `get-f0rise234` takes the times labelled `F0_min` and `F0_peak2` and extracts the intervening CT values, runs them through an averager to get the integrated CT over the rise to the second peak. It then extracts the F0 values at the beginning and end of the rise. It echoes to the screen the integrated CT, and the F0 values at the beginning and end of the rise, and the F0 excursion over the rise. (The averaging program that it calls is

`integ_ascii` which is also in that directory.) These shell scripts are tailored to the specific analyses that we wanted to make, but they illustrate the kinds of very simple scripts that can be made to allow quick batch processing of `waves+` labeled data in such complexes of varied signal types.

2.2. Some results. The results of our analyses suggest that pitch range effects are not a uniform phenomenon across the different levels of the prominence hierarchy. That is, we found different physiological patterns for ostensibly the same patterns of F0 peak patterns, depending on which level of prominence relationship we looked at.

For example, looking at the mean F0 values for the second peak across the three different utterance types, we found that there were differences among the different pitch accents. The peak F0 for the second H* of sentence type 4 was substantially lower than the peak for the L*+H of type 2, which in turn was somewhat lower than the peak for the L+H* of type 3. (This difference was illustrated by the two utterances in Figure 2.) The difference between H* and the other two types seems to reflect something like the inherent prominence from the pragmatic meaning of the accent type. (See Pierrehumbert & Hirschberg, 1990, for the accent meanings, and Ayers, 1995, for further evidence confirming that there is a prominence-related F0 difference between H* and L+H*.). The EMG signal reflected this pattern in the F0 peaks. The mean CT activity level integrated over the rise into the peak was lowest for H* and highest for L+H*.

When we looked at the two kinds of mean values for these peaks across the different vocal effort levels, however, the correspondence was different. As expected from Liberman and Pierrehumbert (1984) and our own earlier results for L tones, we found that the F0 values for these H tones rose substantially for all three accent types in going from soft to normal to loud voice. However, CT activity level did not increase in the same way. The mean levels for soft voice was somewhat lower than those for normal voice, but this difference was very small by comparison to the difference between H* and L+H*, and there was no comparable difference between normal voice and loud voice.

It was interesting to note that the pattern for the P_0 signal in the earlier recording of this corpus by the same speaker showed the opposite correspondence to the F0 peak values. The mean peak P_0 values associated with the second F0 peaks differed substantially across the three levels of overall vocal prominence, but were virtually the same across the three pitch accent types. This suggests that the inherent tonal prominence of the different accents and the overall prominence of different vocal effort levels are not the same physiologically. The higher peak for a L+H* accent compared to a simple H* seems to be produced by an intentional active manipulation of the vocal fold length, whereas the higher peak for a loud-voiced L+H* compared to a soft-voiced production of the same accent type might be a side effect of forcing more air out of the lungs to increase volume. If this interpretation is correct, we might expect there to be spectral differences (more energy at higher frequency bands) for the raised pitch of louder voice, but not for the higher pitch of a L+H* relative to a H*. Further details of the results of our analyses are reported in Beckman, Erickson, Honda, Hirai, & Niimi (1995), which is included as the second appendix to this report.

3. The "East Warsaw Street" corpus

2.1. Background. This corpus was designed explicitly to replicate Sluijter's results for Dutch. We wanted to look at spectral tilt in a corpus of utterances with stressed versus unstressed syllables, at two levels of prominence contrast. In Sluijter and van Heuven's (1993) data, the effect of prominence on spectral tilt

seemed to be much stronger and more consistent when the prominence contrast was between a nuclear accented syllable and a completely unstressed syllable. It was weaker when the more prominent syllable was heavy, but not accented. However, Sluijter looked only at the low central vowel /a/. Thus, it is possible that the effect is limited only to /a/, which has a high first formant and low second formant, and thus a tremendous concentration of energy in the mid-frequency region. That is, if the unstressed vowel were at all reduced toward schwa, the apparent spectral tilt differences could be an artifact of supralaryngeal adjustments to the filter, and not a laryngeal source differences after all. Also, Sluijter did not control for the pitch level. That is, she did not have the speaker avoid the typical Dutch "hat pattern", which puts an F0 peak on the accented syllable, and low pitch around it. Thus, it is possible that the more reliable difference when the prominent syllable was accented could be an artifact of the contrast between higher and lower F0 in that case.

In our corpus, therefore, we wanted to compare nuclear accented, unaccented heavy, and light syllables with vowels other than [a]. Also, we wanted to look at these syllables in low and high pitched regions of the intonation contour. The full corpus is listed in Figure 4 below, which spans the next two and a half pages. The intended intonation contour is transcribed underneath the text, and the target vowel are underlined. The intended prominence levels and pitch pattern were manipulated orthogonally to the target vowel and consonant contexts. Thus, in the set of utterance types (7) to (12), for the vowel [i] in the [b_d] context, we contrasted the maximal stress of nuclear accent on the lexically stressed first syllable of `beadwork` in (7) and (10) with the lesser prominence of the same heavy syllable in post-nuclear unaccented position in (9) and (11), and also with the lexically unstressed first syllable of `bedazzled` in (8) and (12). We also intended to contrast high pitch on the target syllable for utterance types (7) through (9) with low pitch on the syllable for (10) through (12). We had three speakers read five different randomized lists to get 5 tokens of each sentence type. The audio (and laryngograph signals where available) are stored in directory `/usr/pi/data/SPECTRAL-TILT/OCT-17`.

[v_s] context for [i]

1. (How can I get there by car?)
   You should drive EASTWARD from here.
   H* L- L%
2. (Walking to work is hard at this time of year,
   because the sun shines right in my eyes.)
   Isn't it harder to DRIVE eastward then?
   L* H- H%

3. (Where is the house? Is it far?)
   It's the other side of East WARSAW Street?
   H* L- L%
4. (Which direction should I be going?)
   Should I drive EASTWARD to get there?
   L* H- H%
5. (I know it's on this side of Moscow Boulevard, but..)
   is it this side of East WARSAW Street?
   L* H- H%
6. No, it's the OTHER side of East Warsaw Street.
   H* L- L%

[b__d] context for [i]

7. (Their carpentry's a mess, but ..)
   I hear their BEADWORK is good at least.
   H* L- L%
8. I was simply BEDAZZLED by it.
   H* L- L%
9. (Well, their beadwork may be bad, but...)
   have you seen JOEL'S beadwork?
   L* H- H%
10. Have you seen their BEADWORK yet?
    L* H- H%
11. (Why, is their beadwork good?)
    No, they're TERRIBLE at beadwork.
    H* L- L%
12. (How did they react to his new picture?)
    Did they all seem BEDAZZLED by it?
    L* H- H%

[b__b] context for [æ]

13. (Did you hear about John's latest article?)
    He had a story in "The BABOON Report".
    H* L- L%
14. (John's recorded 100 children over the last year.)
    He's studying their BABBLE and such.
    H* L- L%
15. (You're reading the "Babel Mountain Story"?!)
    Wasn't it the "TOWER of Babel Story"?
    L* H- H%
16. (I know you've read about the flood in the Epic of Gilgemash, but)
    Do you know it's also got the Tower of BABEL story?
    L* H- H%
17. (What books by him have you read? For example,)
    Have you ever read "The BABOON Report"?
    L* H- H%
18. ("B" "A" "B" "E" "L")
    No, THAT's not how "babble" is spelled.
    H* L- L%

[z__d] context for [æ]

19. (I know he doesn't like adjectives in these reports, but..)
    do you think we could use ADVERBS in them?
    L* H- H%
20. (I'm sure we don't want to know what Jay DID, but..)
    shouldn't we ask what his ADVICE would be?
    L* H- H%
21. (I know John would say "fortunately" here,)
    But JOAN doesn't use adverbs like this.
    H* L- L%
22. (We're writing for Joan, so...)
    we'd better not use ADVERBS in this.
    H* L- L%
23. (I don't want to copy what he did, but..)
    I'd like to get his ADVICE about it.
    H* L- L%

24. (Okay, so now we know Joan's and Mary's opinions, but..)
Could you tell me what's JAY's advice about this?
L* H- H%

[b__k] context for [u]

25. (Let me tell you about his background...)
He's originally from BUCHWALD.
H* L- L%

26. (Try this wine.)
You'll appreciate its BOUQUET.
H* L- L%

27. (What a fascinating story from those times...)
Were you actually LIVING in Buchwald?
L* H- H%

28. (Is his whole family from there?)
No, only his MOTHER's from Buchwald.
H* L- L%

29. (Tell me about your trip. For example, ...)
are you planning to stop in BUCHWALD.
L* H- H%

30. (What's wrong with the wine...)
Do you like it's BOUQUET at least?
L* H- H%

[t__t] context for [u]

31. (Tell me why you don't like my vacation plans.)
The trip starts at UTRECHT.
H* L- L%

32. (Try this novel...)
You'll find it OUTRE.
H* L- L%

33. (You keep talking about prices just from Utrecht, so ..)
does the trip START at Utrecht?
L* H- H%

34. (You keep talking about how to change trains in Utrecht, but ...)
my trip STARTS at Utrecht.
H* L- L%

35. (Tell me about your trip. For example, ...)
does the trip start at UTRECHT?
L* H- H%

36. (Would you recommend this for Jay? Or..)
do you think he'll think it OUTRE?
L* H- H%

3.2. The database. The structure of the database stored in directory `/usr/pi/data/SPECTRAL-TILT/OCT17-94` is as follows. There are five subdirectories: `KL`, `NC`, `MB`, `notes`, and `programs`.

`KL NC MB`: These first three directories contain the speech files for the utterances themselves, and each is named after the speaker who produced the utterances in that subdirectory. The data files are called by base names such as `kl034`, where `kl` is the initials of speaker KL, and `034` is the number of the utterance in the corpus

list. Utterances for which there is a usable electroglottograph (EGG) signal have associated data files of three types:

`kl034.d`	the original stereo file extracted from the super long digitized file for the speaker's reading
`kl034.egg.d`	the demuxed mono file for the EGG data
`kl034.sp.for`	the demuxed mono file for the speech data

Utterances recorded after battery power on the EGG had run low have only the demuxed speech data. Each utterance also has an associated F0 file produced by get_f0 (e.g. `kl034.f0`), and a label file (e.g. `kl034.lab`) where Y. Ohta has marked the beginning (`V_start`) and end (`V_end`) of the target vowel for us. The shell `show-f0` in each directory is a `waves+` script for displaying the data and labels, with the F0 range set for the speaker's approximate range. See Figure 5 for sample displays of two utterances by `KL`, where the target vowel is nuclear-accented on low pitch (top half of figure) versus unaccented in the high pitched post-nuclear tail (bottom half of figure).

`notes`: This directory contains various files, including the list of the five repetitions of the 36 sentence types in the order in which they were read. Other important files here are `spectral.tilt.key`, which lists the token numbers for each utterance type after a three-digit code stating the intended prominence level, pitch level, and target vowel, and the three files `spectral.tilt.KLnotes`, `spectral.tilt.NCnotes`, and `spectral.tilt.MBnotes`, which contain my notes about the actual intonation contours produced around the target vowel in each of KL's, NC's, and MB's tokens. Since the speakers did not always produce the intonation contour that we intended, these notes are important for interpreting the results of the spectral tilt analysis.

`programs`: This directory contains source code for two c programs and two shell scripts that Nick Campbell wrote to do the kinds of spectral tilt calculations that we used in Campbell & Beckman (1995). The files are:

`stats.loop`
A shell to run the spectral tilt extraction shell over all of the speech files in a database.

`stats.sh`
The actual spectral tilt extraction shell script. This computes an f0 file (if there isn't one), calculates an FFT to send to the `pitchproc` program, merges the output of `pitchproc` with the f0 file, and then calls `hmm2stats` to calculate means over labelled intervals

`pitchproc.c`
Calculates (1) the intensity at the fundamental, (2) the ratio between it and energy at the second harmonic, (3) total energy in a midfrequency band (scaled to average energy over all frequencies). This now calculates value (3) over the band between 2kHz and 4kHz, if the input speech file is sampled at 16kHz. These lines are clearly identified in a comment, so you can adjust the value for other sampling rates or other frequency bands.

`hmm2stats.c`
calculates the average values for all of the channels in the merged f0 file over an interval defined by a label file

Note that these utilities use ESPS routines, so you need to have an ESPS license checked out to run them.

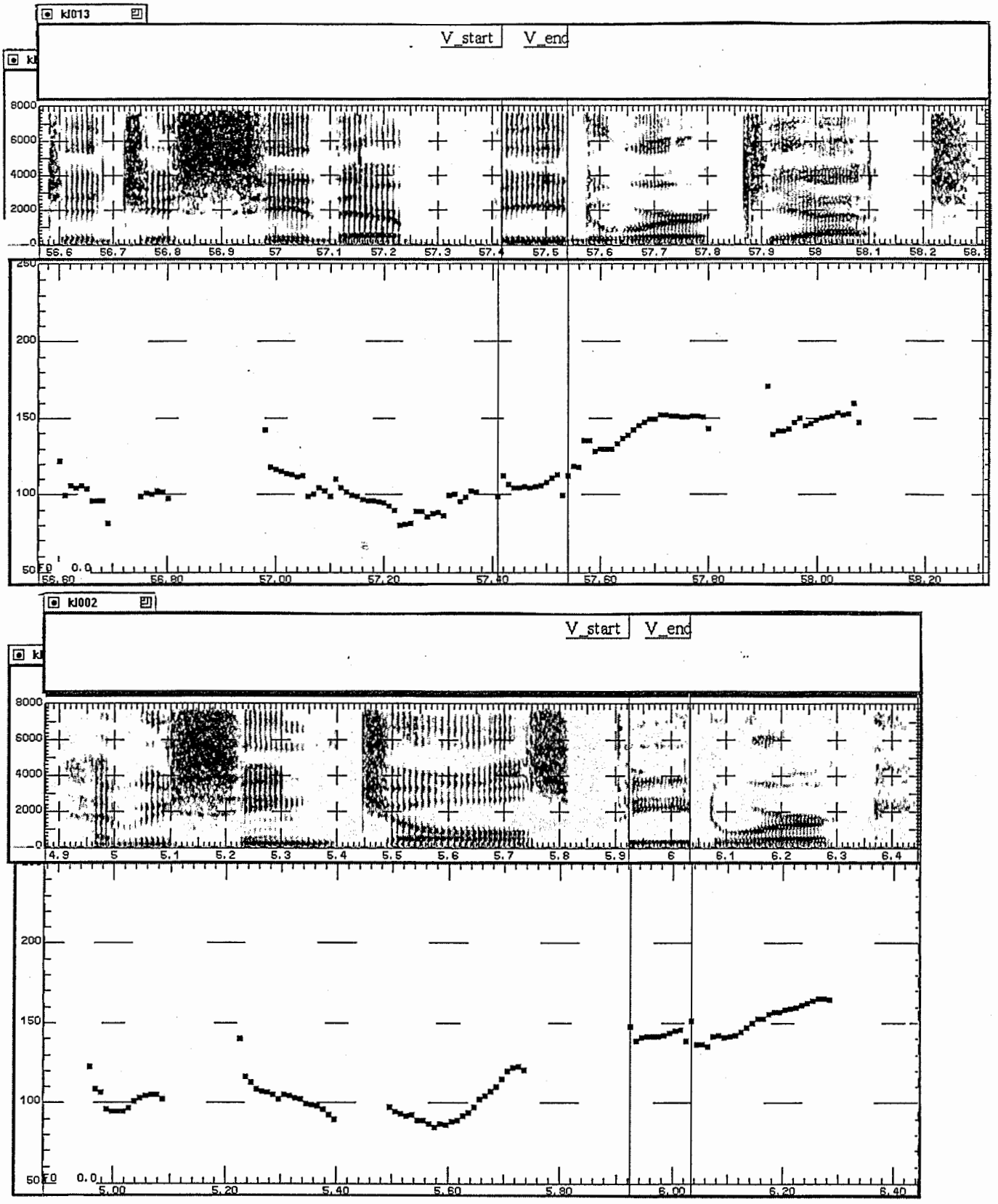


Fig. 5. Sample displays of utterance k1013 "Have you seen their BEADWORK yet?" (top set of window) and utterance k1002 "Have you seen JOEL'S beadwork?" (bottom set of windows) .

3.3. Some results. Initial results of our analysis were presented at the Spring, 1995, meeting of the Acoustical Society of Japan. The two-page abstract from the proceedings is attached in Appendix C. When we compared the energy in different parts of the mean spectrum for accented vowels compared to the same frequency bands for pre-nuclear lexically unstressed and post-nuclear unaccented vowels, we found larger, more significant differences at higher frequency bands. This was true for the corpus as a whole, and it was also true for each vowel type averaged separately. These results are promising. Particularly, the pattern of differences in the comparisons involving [i] supports Sluijter's interpretation of the difference she found for [a]. However, because we recorded only five repetitions of each utterance type per speaker, and because the speakers did not always produce the targeted intonation contour, we did not have enough tokens in each cell to do a full ANOVA to disentangle the three factors varied (namely PITCH, SPEAKER, and VOWEL TYPE). Thus the results must remain inconclusive.

To overcome these limitations, we have devised a new corpus (described in Figure 6), which was more successful at eliciting the intended intonation types. We have recorded audio and laryngograph for 10 repetitions of each utterance types produced by four speakers as of May 25, 1995. Ms. Ohta is currently digitizing and labeling the target vowels in these utterances and they will be stored in directory `/usr/pi/data/SPECTRAL-TILT/MAY25-95`.

The corpus gives an (almost completely) orthogonal variation among three factors:

(1) Target vowel [æ] vs [i] vs [u] in B*á*ddle vs B*éa*dle vs B*óo*dle, and B*a*dd-Éllis vs B*e*de-Éllis vs B*oo*de-Éllis.

(2) Primary lexical stress placement on target vowel versus on following syllable in B*á*ddle vs B*a*dd-Éllis, etc.

(3) Nuclear-accented vs postnuclear unaccented in:

```
He's interviewed ALL of the men from that gang.
He's done Tony Luciano,
          Jonathon Báddle, Nathanial Jackson, ...
                   H*   H-

He's written books on ALL of the famous Baddles.
He's done Matthew Baddle,
             Jónathon Baddle, Miriam Baddle, ...
              H*       H-
```

(4) High pitched vs low-pitched postnuclear in the utterance type above vs :

```
No, it's not JONATHON Baddle I interviewed, but
                                his brother, Matthew.
              H*       L-
```

Fig. 6. New corpus for spectral tilt analysis.

4. The "Mary Anaheim" corpus

4.1. Background. The purposes of this corpus were many. First, we wanted to re-examine as many of the questions as possible that had been partially addressed by the first two corpora. That is, we wanted to look at pitch range in phrases produced in several degrees of overall vocal effort, and at spectral tilt in vowels with varying degrees of prominence, ranging from completely unstressed to nuclear accented. At the same time, we wanted to look at the relationships among

fundamental frequency, subglottal pressure, and EMG activity level in as many laryngeal muscles as we could record simultaneously, for a wide variety of intonation types. Part of the motivation for this was to be able to do utterance-by-utterance correlations between the two different types of physiological measure, in order to be more confident of our interpretation of the different pitch raising mechanisms involved in higher overall vocal effort versus inherently higher accent peaks. (This we could not do with the data for the first corpus, because the subglottal pressure and EMG were obtained in separate recording sessions.) Another part of the motivation, however, was to try to build some kind of computational mapping between the EMG signals and the ToBI labelled intonation pattern, analogous to the statistical mappings between lingual muscle activity and vowel formants that Dr. Honda has been working on in collaboration with various people both here at ATR-HIP and elsewhere (see, e.g., Kakita, Fujimura, & Honda, 1985; Honda, K.; Honda & Maeda, 1995).

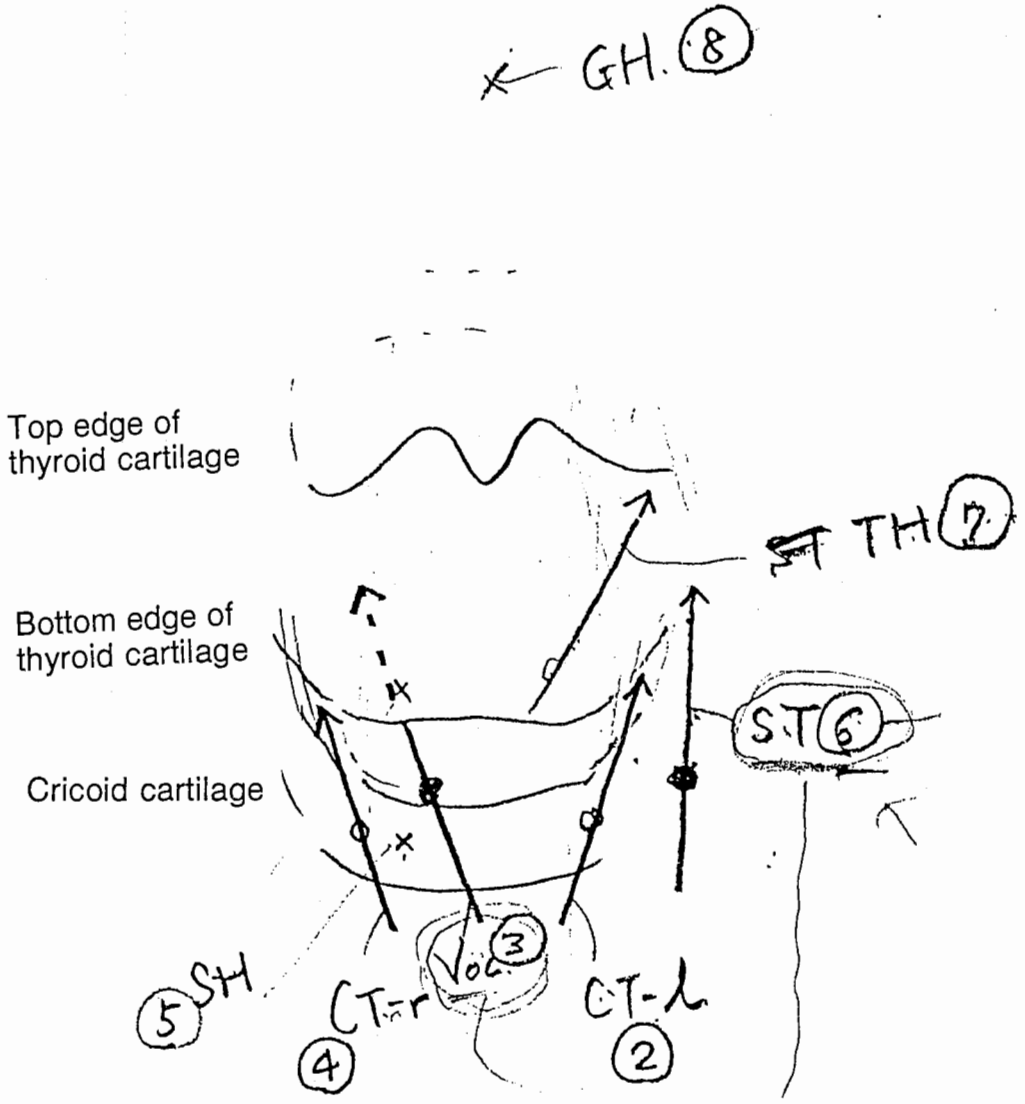


Fig. 7. Dr. Honda's drawing of electrode insertion sites for the Mary Anaheim corpus. It is a frontal view, showing the presumed shapes of the thyroid and cricoid cartilages, with insertion angles and channel order.

For this corpus, we decided to record EMG signals from as many different intrinsic and extrinsic laryngeal muscles as we could without inserting the electrodes periorally. (This was also the method used in the first corpus.) We also decided to get, for at least a subset of the utterances, simultaneous subglottal pressure. Level of EMG activity was recorded from hooked-wire electrodes inserted subcutaneously through the skin of the neck, as in the first database. (The method is described in detail in Hirose, Gay, and Shome, 1971.) Figure 7 is the drawing of the insertion sites that Dr. Honda made at the time of the recording. Subgottal pressure was measured as in the earlier experiment described in Erickson, Honda, Hirai, Beckman, & Niimi (1994). We used a transducer designed to measure arterial blood pressure, mounted at the tip of a catheter, which we could then insert through the nose and over the velopharyngeal port to pass down the pharynx into the subglottal tracheal tube. Kevin Lenzo video-taped the entire recording session, and Nick Campbell has the videotapes, which you are welcome to look at if you are not squeamish.

The corpus was a set of three women's names produced in the intonation patterns described in Figure 8. The choice of names by itself lets us contrast

(1) hat pattern

```
Mary Anaheim
 H*  H* L- L%

Marianna Heim      or          Marianna Heim
     H*    H* L- L%             H*       H* L- L%

Marie Annapolis
     H*   H* L- L%
```

(2) contrastive accent with late nucleus

```
Mary Anaheim
     L+H* L- L%
```

(3) contrastive accent with early nucleus

```
Mary Anaheim
L+H*     L- L%
```

(4) uncertainty contour

```
Mary Anaheim
     L*+H L- H%
```

(5) calling contour

```
Mary Anaheim        or          Mary Anaheim
 H*  H* !H- L%                      L+H* !H- L%
```

(6) surprise-redundancy contour

```
Mary Anaheim
 L*  H*  L- L%
```

(7) yes-no question contour with late nucleus

```
Mary Anaheim
 L*  L*  H- H%
```

(8) yes-no question contour with early nucleus

```
Mary Anaheim
 L*  H-   H%
```

Fig. 8. Intended intonation types for the Mary Anaheim corpus.

several different levels of prominence on several different vowel types. For example, the high front vowel [i] is common to the second syllable of all three of the names. In the names `Mary Anaheim` and `Marianna Heim`, the vowel is completely unstressed and liable to reduction, whereas in `Marie Annapolis` it has primary lexical stress and is always the full tense vowel. When we consider the intonation pattern, then, we get several more contrasts, because the primary lexical stress means that this syllable in `Marie Annapolis` will take the nuclear accent in contours (3) and (8) "with early nucleus", but will be prenuclear and perhaps unaccented in the corresponding contours "with late nucleus" in types (2) and (7). Similarly, the low front vowel [æ] is common to the names `Mary Anaheim` and `Marianna Heim`, but its stress status differs. In `Mary Anaheim` it has primary lexical stress and will bear the nuclear accent in most of the intonation contours, whereas in `Marianna Heim`, it will typically be prenuclear. The [ai] diphthong in the last syllable in these two names shows an analogous contrast between having the primary lexical stress in `Heim`, and being always postnuclear and therefore unaccented in `Anaheim`.

As for the first database, the subject was me. I produced 15 tokens of each type at two different overall vocal effort levels — normal and loud. Five of the repetitions were before the catheter was inserted to measure subglottal pressure, five were produced with the catheter, and another five were produced after the catheter had to be removed because the topical anesthetic was beginning to wear off. During the middle five repetitions, one of the CT EMG signals was unhooked from the FM recorder, and replaced by the signal from the pressure transducer. The digitized and processed signals for these repetitions (and for some snippets of spontaneous speech produced during the recording session) then constitute the database, stored in directory `/usr/pi/data/EMG-P0-93`.

4.2. The database. The structure of the database of utterances stored in directory `/usr/pi/data/EMG-P0-93` is as follows. There are four types of sets of data files, for: (1) target utterances with 7 different EMG signal traces, including both right and left CT signals, (2) target utterances with subglottal pressure instead of one of the CT signals, and (3-4) five spontaneous utterances such as "The next one is gonna have to be the last, because I'm starting to feel it." (where "it" refers to the catheter with the pressure transducer). The last constitutes two types, because the speech can have two CT traces or only one CT trace and a P_0 trace, depending on when in the recording session I produced it. The basenames for the spontaneous utterances are of the type `Spontaneous3` (for number 3 of the five spontaneous speech utterances). The basenames for the target utterances are of the type `Be1-2L`, `Du3-8N`, and `Af4-5La`. In these names, the `Be`, `Du`, and `Af` refer to the sections of the recording session before, during, and after the time where I was speaking with the pressure transducer hanging down between my glottis, respectively; the first digit (the one before the hyphen) refers to the repetition number; the second digit refers to the target utterance type, as listed in Figure 8; and the capital letter refers to whether the utterance was produced in `Normal` or `Loud` voice. (The occasional "a" at the end of some basenames merely means that I had to try a second digitizing batch command with the untested fast sampling rate on the multichannel DAT recorder before I got the entire utterance.)

Associated with each database are a set of up to 16 or 15 files, depending on whether the utterance was produced during the middle third of the recording session, when the subglottal pressure signal replaced the second CT trace. Six of these files are ASCII files containing the text of the utterances, or labels produced by the aligner program, or by ToBI labeling which I had only just begun to do when I left ATR-ITL to return to Ohio State University. The ToBI labeling is not finished,

and so some of the utterances have no associated .tones or .misc files. (They all have .breaks files, since I created ToBI break index place holders by copying the .words files produced by the Align program.) The data in the files for the 6 or 7 EMG signals have all been downsampled, rectified, and smoothed, in the same way that I smoothed the EMG signals in the first database. (In this database, the raw data were stored in temporary files which were deleted after being processed.) There are two types of subglottal pressure datafiles, one for the raw data, and the other for downsampled and smoothed data. (The smoothing eliminates the high-frequency fluctuations due the periodic opening and closing of the glottis during voiced portions of the utterances.) The file extensions are shown in the following list of names:

Be1-2L.sp.d	the audio file
Be1-2L.f0	the F0 file calculated with get_f0
Be1-2L.ct1.d	EMG signal for the left cricothyroid muscle
Be1-2L.ct2.d	and for the right cricothyroid muscle
Be1-2L.ta.d	EMG signal for the thyroarytenoid
Be1-2L.gh.d	EMG signal for the geniohyiod
Be1-2L.th.d	EMG signal for the thyrohyoid
Be1-2L.st.d	EMG signal for the sternothyroid
Be1-2L.sh.d	EMG signal for the sternohyoid
Du1-2L.p0.d	unsmoothed subglottal pressure
Du1-2L.p0.av.d	downsampled and smoothed subglottal pressure
Be1-2L.txt	ASCII of text of utterance
Be1-2L.words	xlabel file for words
Be1-2L.phones	for phones
Be1-2L.breaks	for break indices
Be1-2L.tones	for tones
Be1-2L.misc	for miscellaneous comments

The directory also contains some shell scripts for displaying various combinations of data in xwaves. They include:

show-emg-all	complete display of all traces for utterances with no P_0 trace.
show-emg	nicely layered display for same.
show-p0	display of traces for utterances with a P_0 trace.

Figures 9 and 10 show sample displays made with the two relevant scripts for utterance Be1-2L, and Figure 11 shows a sample display for the relevant script for Du1-2L. The file notes-all contains notes on the utterances that I have already done the ToBI labelling.

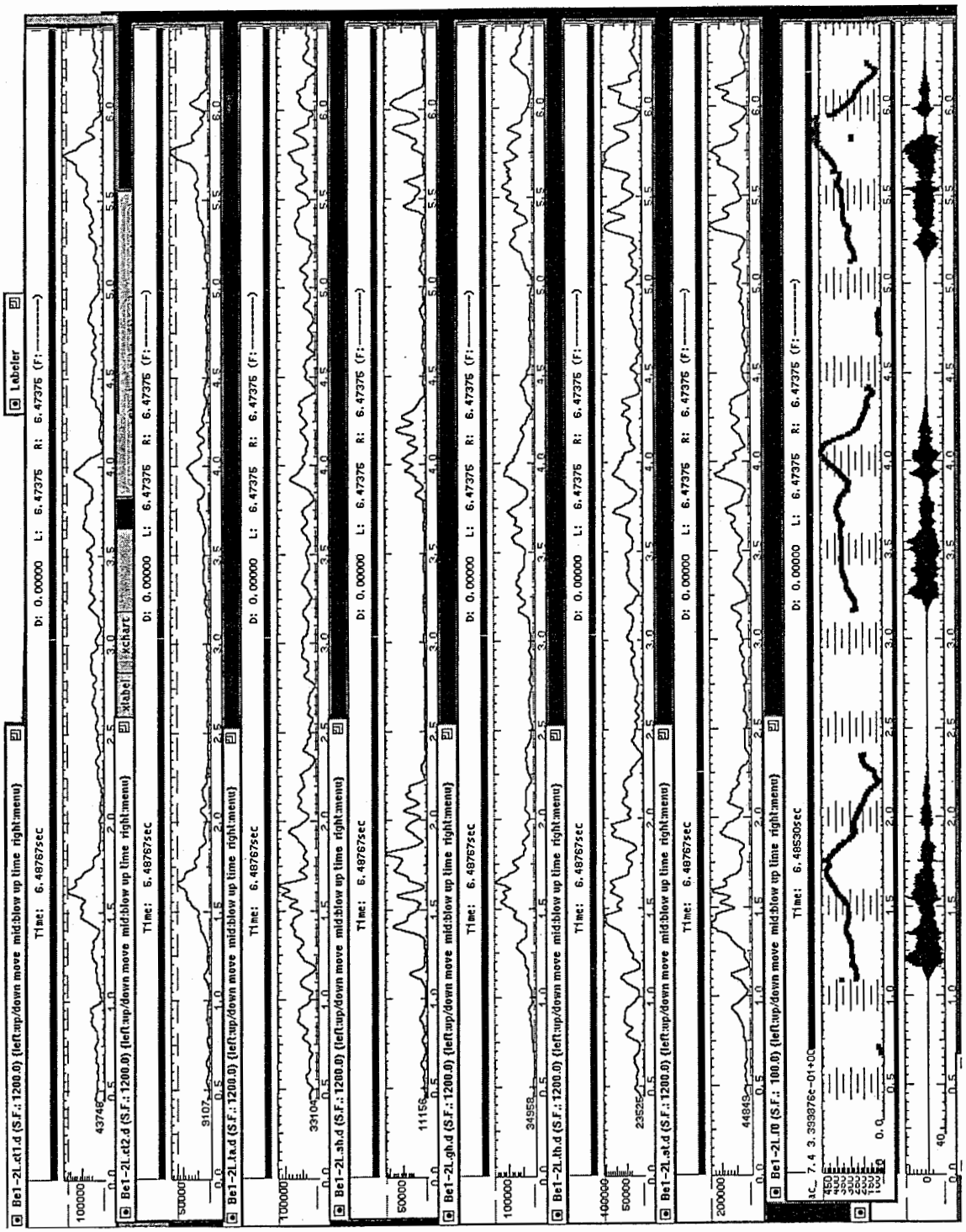


Fig. 9. Sample display from `show-emg-all`.

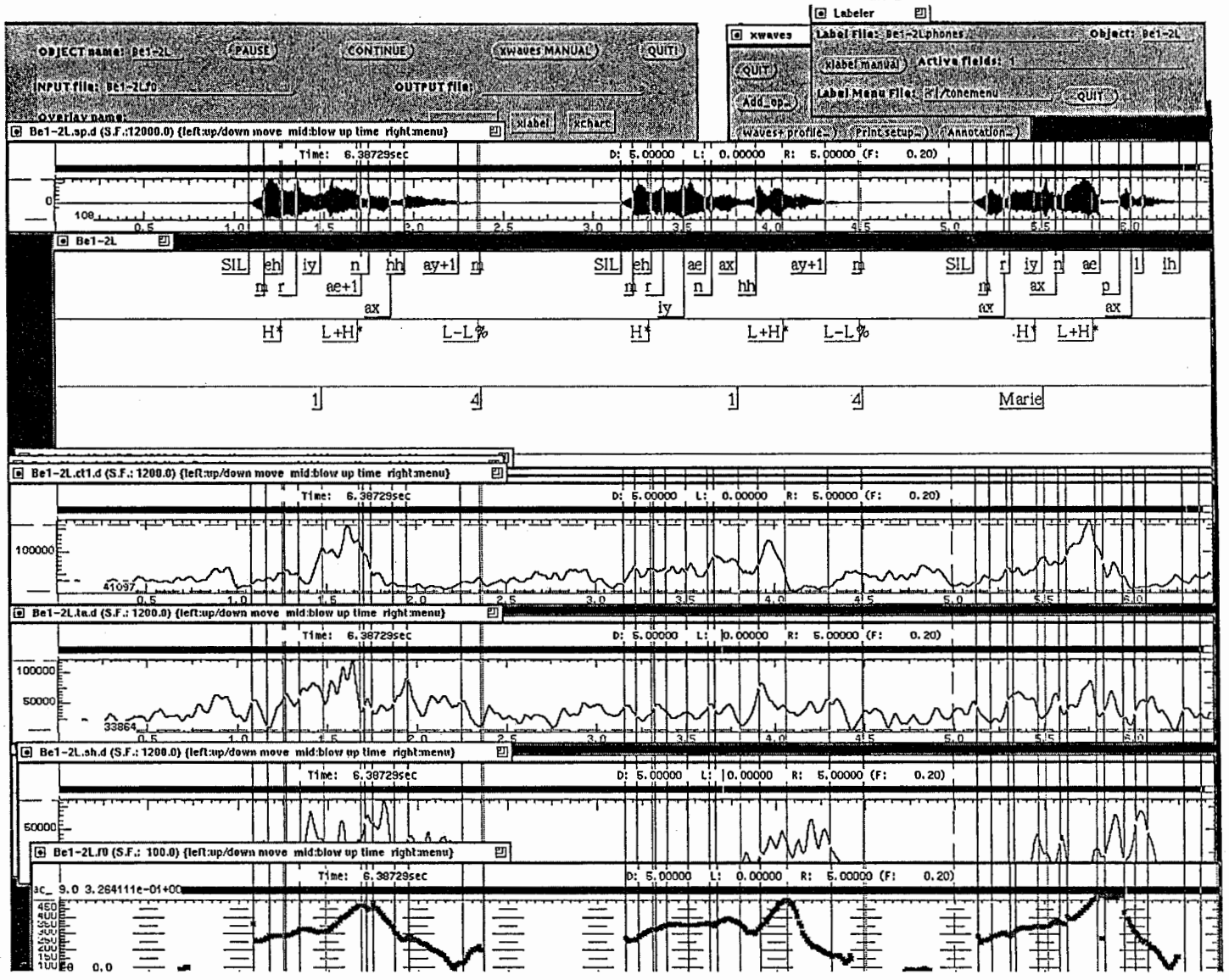


Fig. 10. Sample display from show-emg.

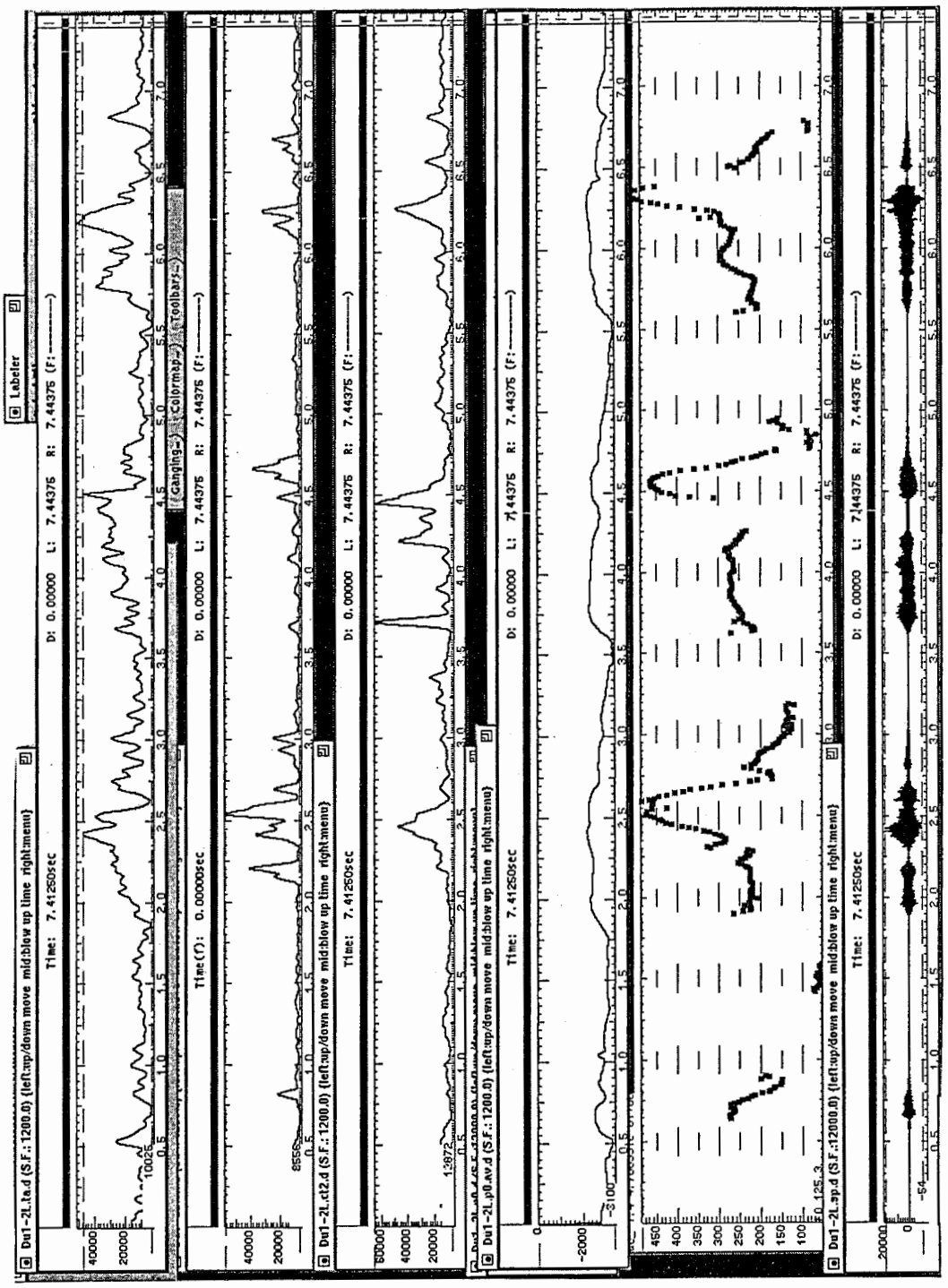


Fig. 11. Sample display from `show-P0`.

4.3. Some results. Since processing of the database is not yet complete, we have no quantitative results to report on the spectral-tilt and physiological analyses as of this date. However, some qualitative observations suggest that these data will be useful for many more things even than the ambitious analyses for which we designed the corpus. Most notably, since the recordings include records of the thyroarytenoid as well as the cricothyroid, we may be able to use the data to examine the physiological correlates of the many different kinds of acoustic phenomena that are heard as glottal stop. (The thyroarytenoid is the pair of muscles that constitute the main muscle fiber in the vocal folds themselves.) The database should provide many examples of several of the kinds of phenomena that are associated with the percept of glottal stop. For example, since the surname `Anaheim` begins with a stressed syllable and no onset consonant, we might expect to see many occurrences of glottal stop demarcating the stressed syllable onset, particularly in the many intonation types which put nuclear accent on that syllable. Also, the database includes many repairs. Labellers often perceive a glottal stop sharply cutting off phonation if the repaired item is interrupted in the middle of a word.

The remaining figures give some examples of glottal-stop related phenomena that I have noticed in just the few utterances which I managed to ToBI label before leaving ATR-ILT. Figure 12 shows a classic glottal stop, marked by `q` in the phone label window, at the beginning of the nuclear-accented vowel initial syllable. This particular example is unusual in that the stressed vowel is foot-initial, but word-medial. Figure 13 gives another very common realization of this kind of foot-initial "glottal stop" — creaky voice. The `q` in the phone labels marks the offset of the creaky voice portion. This example also shows that the creaky phonation need not be limited to the beginning part of the stressed vowel, since here the main portion of creak is manifest over the end of the [i] in the preceding `Mary`. Both of these figures show a substantial increase in the thyroarytenoid activity level, with no corresponding increase in the cricothyroid signal. Previous work on laryngeal activity associated with raising and lowering pitch shows that the two sets of muscles are highly correlated, in a way that suggests that they contract together to raise pitch. (See the review and discussion in Titze, 1994, chapter 8.) Perhaps the increase in thyroarytenoid with no corresponding increase in cricothyroid here reflects tensing of the vocal folds for the glottal stop. It will be interesting to see whether this interpretation can be sustained in a more detailed examination of the relationship between TA and CT throughout this database.

Figure 14 shows a third example where both I and Nick Campbell clearly perceived this kind of foot-initial glottal stop. Here, unlike in Figures 12 and 13, there is no indication of a classic glottal stop or creak in the spectrogram. The percept seems to be related to the "segmental" effect of a sudden sharp dip in the fundamental frequency. It is interesting to note that this utterance also shows the increase in TA with no corresponding increase in CT just before the perceived glottal stop.

Figure 15 shows an example of perceived glottal stop when phonation is abruptly cut off before a repair. Here, by contrast, there is no increase in TA activity level. While there is much work left to be done on this database, whatever results we find are bound to be interesting and potentially applicable to our speech synthesis and recognition efforts.

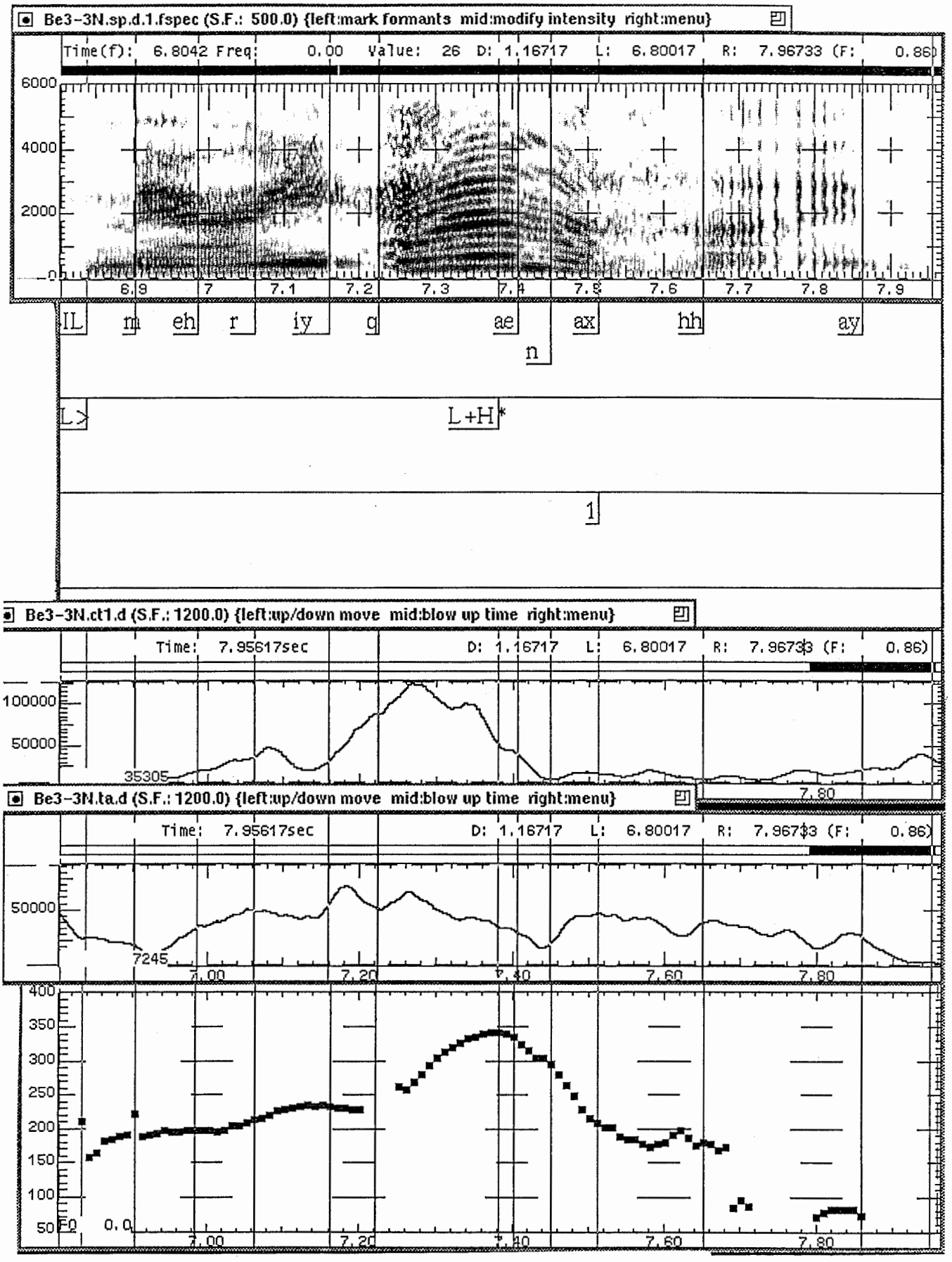


Fig. 12. Example of a classic foot-initial glottal stop.

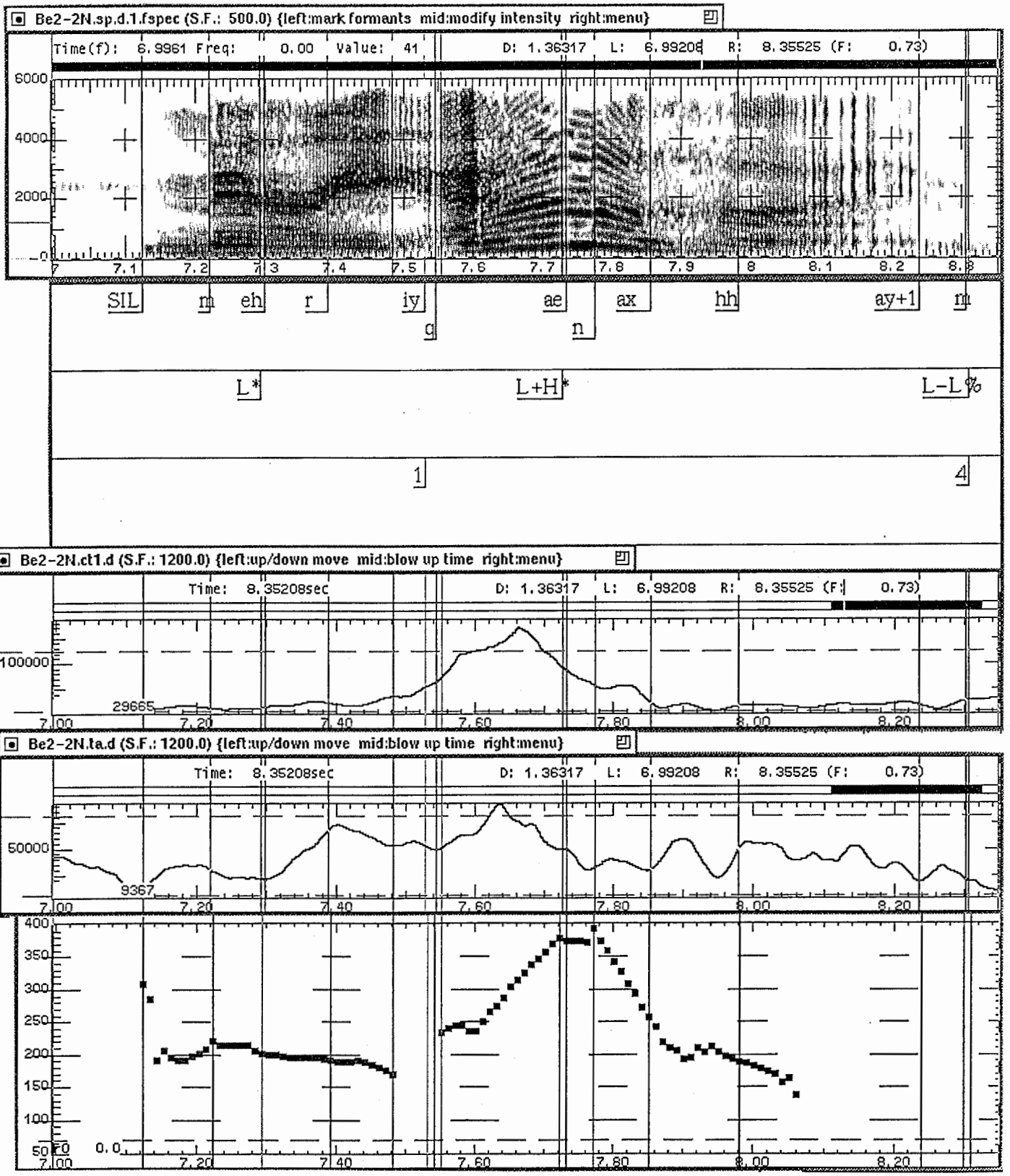


Fig. 13. Example of the creaky voice realization of foot-initial glottal stop.

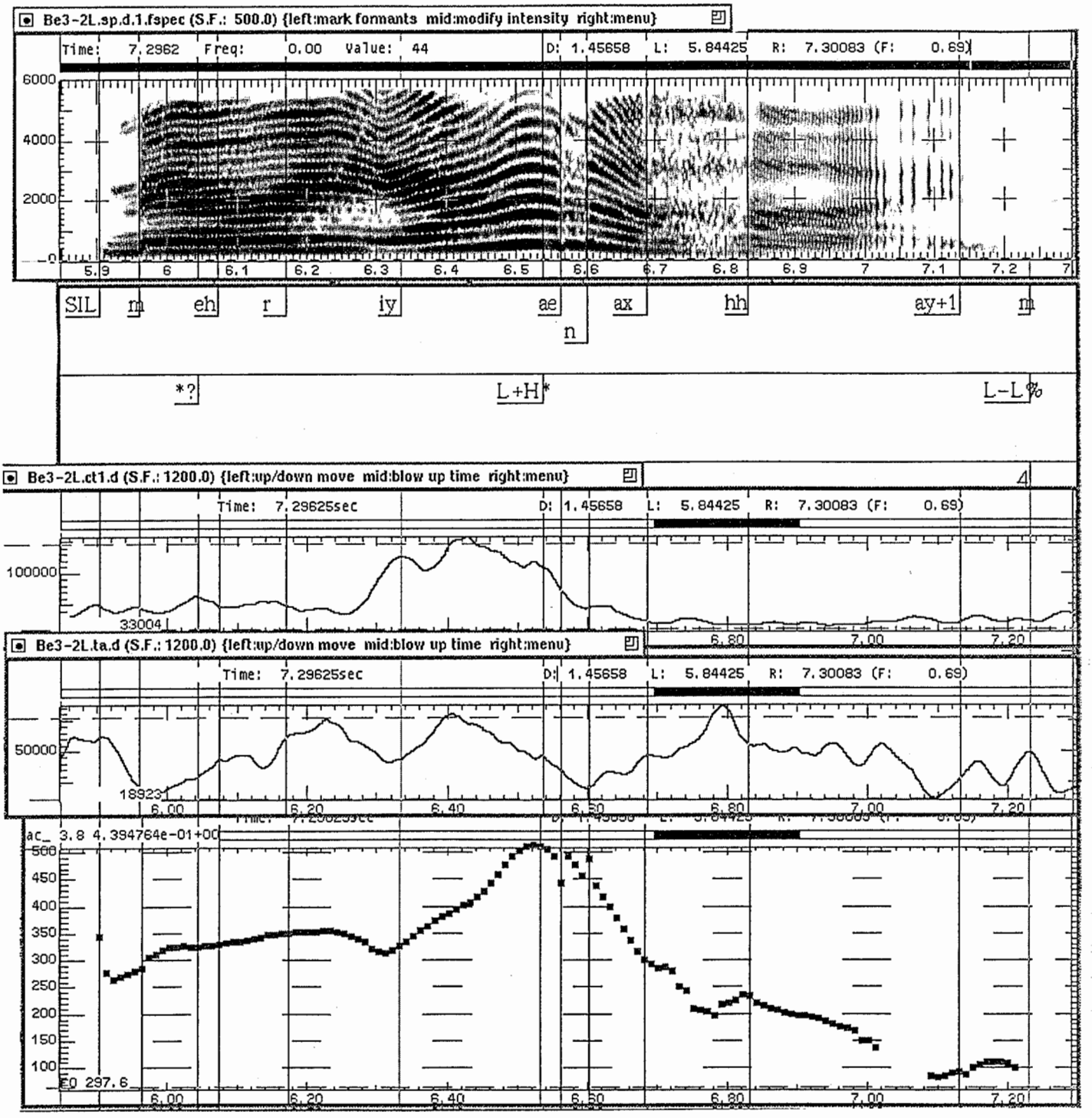


Fig. 14. A perceived glottal stop with no apparent creak.

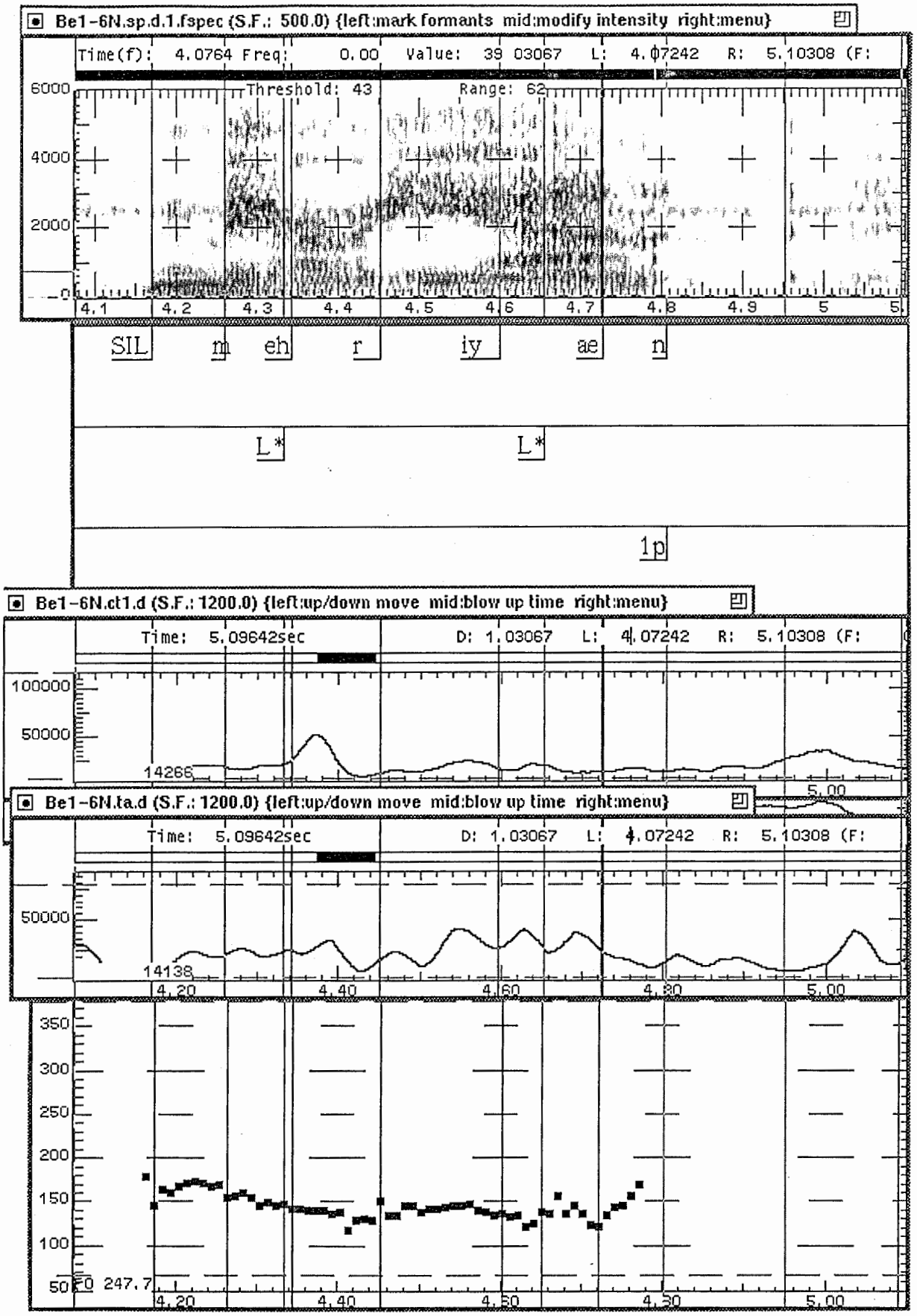


Fig. 15. Perceived glottal stop before repair.

References

Avesani, C., & Vayra, M. (1988). "Discorso, segmenti di discorse e un'ipotesi sull'intonazione," *Att del Convegno Internazionale "Sull'Interpunzione"*, Florence.

Ayers, G. M. (1995). "Nuclear accent types and prominence: some psycholinguistic experiments," To appear in *Proceedings of the 13th International Congress of Phonetic Sciences.*

Beckman, M. E. (1986). *Stress and Non-Stress Accent.* Dordrecht: Foris Publications.

Beckman, M. E., & Ayers, G. M. (1994). *Guidelines to ToBI Labelling, ver. 2.* Tutorial description and accompanying labelled speech examples, Ohio State University. [Available here at ATR-ITL in directory /usr/pi/data/ToBI.]

Beckman, M. E., & Edwards, J. (1992). "Intonational categories and the articulatory control of duration," In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka, eds., *Speech Perception, Production and Linguistic Structure*, pp. 457-463. Tokyo: OHM Publishing Co.

Beckman, M. E., & Pierrehumbert, J. (1986). "Intonational structure in English and Japanese," *Phonology Yearbook*, 3, 255-309.

Beckman, M., Erickson, D., Honda, K., Hirai, H., & Niimi, S. (1995). "Physiological correlates of global and local pitch range variation in the production of high tones in English," To appear in *Proceedings of the 13th International Congress of Phonetic Sciences.*

Campbell, W. N. (in press). "Prosody and the selection of units for concatenation synthesis," To appear in J. P. H. van Santen, ed., *Progress in Speech Synthesis.* New York: Springer-Verlag.

Erickson, D., Honda, K., Hirai, H., & Beckman, M. (1993). "The production of low tones in English intonation," *ATR Technical Report TR-H-023.*

Erickson, D., Honda, K., Hirai, H., & Beckman, M. (1995). "The production of low tones in English intonation," *Journal of Phonetics*, 23, 179-188.

Erickson, D., Honda, K., Hirai, H., Beckman, M., & Niimi, S. (1994). "Global pitch range and the production of low tones in English intonation," *Proceedings of the 1994 International Conference on Spoken Language Processing*, pp. 651-654.

Fry, D. B. (1955). "Duration and intensity as physical correlates of linguistic stress," *Journal of the Acoustical Society of America*, 27, 765-768.

Fry, D. B. (1958). "Experiments in the perception of stress," *Language and Speech*, 1, 126-152.

Fujisaki, H., & Sudo, H. (1971). "Synthesis by rule of prosodic features of connected Japanese," *Proceedings of the 7th International Congress of Acoustics*, vol. 3, pp. 133-136.

Fujisaki, H., Ohno, S., Masafumi, O., Sakata, M., & Hirose, K. (1994). "Prosodic characteristics of a spoken dialogue for information query," *Proceedings of the 1994 International Conference on Spoken Language Processing*, pp. 1103-1106.

Gauffin, J., & Sundberg, J. (1989). "Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech and Hearing Research*, 32, 556-565.

Higuchi, N., Hirai, T., & Sagisaka, Y. (1994). "Effect of speaking style on parameters of fundamental frequency contour," In *Proceedings of the 2nd ESCA/IEEE Workshop in Speech Synthesis*, pp. 135-138.

Hirose, T., Gay, M., & Shome, M. (1971). "Electrode insertion techniques for laryngeal electromyography," *Journal of the Acoustical Society of America*, 50, 1449-150.

Honda, K. (1991). "A statistical analysis of tongue muscle activity and vowel formant frequencies," *Journal of the Acoustical Society of America*, 90, 2310. [Abstract.]

Honda, K., Hirai, H., & Kusakawa, N. (1993). "Modeling vocal tract organs based on MRI and EMG observations and its implication on brain function," *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 27, 37-49.

International Standard ISO 532-1975(E) (1975). "Acoustics — Method for calculating loudness level."

Kakita, Y., Fujimura, O., & Honda, K. (1985). "Computation of mapping from muscular contraction to formant patterns in vowel space," In V. Fromkin, ed., *Phonetic Linguistics*, pp. 133-144. New York: Academic Press.

Ladd, D. R. (1980). *The Structure of Intonational Meaning*. Bloomington, IN: Indiana University Press.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, M., & Pierrehumbert, J. (1984). "Intonational invariance under changes in pitch range and length," In M. Aranoff & R. T. Oehrle, eds., *Language Sound Structure: Studies in Phonology Presented to Morris Halle*, pp. 157-233. Cambridge, MA: MIT Press.

Maeda, S., & Honda, K. (1994). "From EMG to formant patterns of vowels: the implications of vowel spaces," *Phonetica,* 51, 17-29.

Maekawa, K. (1991). "Tookyoo hoogen gimonbun no intoneesyon," *Nihon Onsei Gakkai Zenkoku-Taikai Kenkyuu-Happyoo-Ronsyuu*, pp. 42-47.

Maekawa, K. (1995). "Effects of focus on vowel formant frequencies in Japanese," *ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*.

Nakajima, S., & Tsukada, H. (1995). "Prosodic features of utterances in task-oriented dialogues," *ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*.

Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, UK: MIT Press.

Pierrehumbert, J. & Hirschberg, J. (1990). "The meaning of intonation contours in the interpretation of discourse." In P. R. Cohen, J. Morgan, & M. E. Pollack, eds., *Intentions in Communication,* pp. 271-311. Cambridge, MA: MIT Press.

Pitrelli, J., Beckman, M. E. & Hirschberg, J. (1994). "Evaluation of prosodic transcription labeling reliability in the ToBI framework," *Proceedings of the*

1994 International Conference on Spoken Language Processing, pp. 123-126.

Sawashima, M., Hirose, H., Yoshioka, H., Horiguchi, S., & Kiritani, S. (1983). "Interaction between jaw movement and vocal pitch control," In A. Cohen & M. van den Broecke, eds., *Abstracts from the 10th International Congress of Phonetic Sciences*, p. 454. Dordrecht: Foris Publications.

Silverman, K. (1977). *The Structure and Processing of Fundamental Frequency Contours*. Doctoral thesis, University of Cambridge.

Sluijter, A. M. C., & van Heuven, V. J. (1993). "Perceptual cues of linguistic stress: intensity revisited," *Proceedings of an ESCA Workshop on Prosody, Lund University Working Papers*, No. 41, pp. 246-249.

Sluijter, A. M. C. (1994). "Vocal effort as a cue for linguistic stress," *Journal of the Acoustical Society of America*, 95, 2873. [Abstract.]

Swerts, M., & Geluykens, R. (1994). "Prosody as a marker of information flow in spoken discourse," *Language and Speech*, 37, 21-43.

Takeda, S., & Ichikawa, A. (1990). "Analysis of prosodic features of prominence in spoken Japanese sentences," *Proceedings of the 1990 International Conference on Spoken Language Processing*, pp. 493-496.

Titze, I. R. (1994). *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice-Hall.

Tsumaki, J. (1994). "Intonational properties of adverbs in Tokyo Japanese," *Proceedings of the 1994 International Conference on Spoken Language Processing*, pp. 1727-1730.

Venditti, J., Jun, S.-A., & Beckman, M. E. (in press). "Prosodic cues to syntactic and other linguistic structures in Japanese, Korean, and English," To appear in J. Morgan & K. Demuth, eds., *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

Yoshida, Y., Honda, K., & Kakita, Y. (1992). "Noninvasive EMG measurement of laryngeal muscles and physiological mechanisms of prosody control," *Densi Zyoohoo Tuusin Gakkai, SP-91-124*, pp. 17-24.