TR-IT-0111

Speech Alignment and Prosodic Transcription

Rachael Serrell
Ohta Yoko
Nick Campbell

1995.4

ABSTRACT This paper examines the process of aligning speech files in terms of words and syllables taken from an utterance of spontaneous or non-spontaneous speech. It is designed to be used as an instruction manual with example files listed for reference. The same example files are used throughout to enable continual reference access to known data. The example files are taken from two databases currently under analysis for their spontaneous speech content. The first, *Sally* is a monologue consisting in its original form of twenty minutes of the speakers' recollections and feelings about Japan. The entire process of files used for aligning *Sally* is referred to in the brackets after each command. The second *emmi*, taken from a different speaker, is a set of dialogues collected from a multi-media experiment where a travel agent advises a client on the various ways to get to his desired destination. The example files given here are taken from the agents' side of the conversation.

1 Abstract

This paper examines the process of aligning speech files in terms of words and syllables taken from an utterance of spontaneous or non-spontaneous speech. It is designed to be used as an instruction manual with example files listed for reference. The same example files are used throughout to enable continual reference access to known data. The example files are taken from two databases currently under analysis for their spontaneous speech content. The first, *Sally* is a monologue consisting in its original form of twenty minutes of the speakers' recollections and feelings about Japan. The entire process of files used for aligning *Sally* is referred to in the brackets after each command. The second *emmi*, taken from a different speaker, is a set of dialogues collected from a multi-media experiment where a travel agent advises a client on the various ways to get to his desired destination. The example files given here are taken from the agents' side of the conversation.

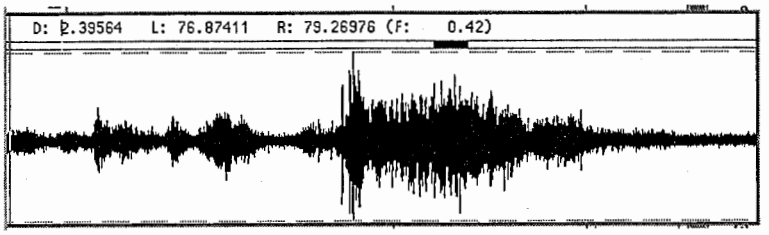
2 Introduction

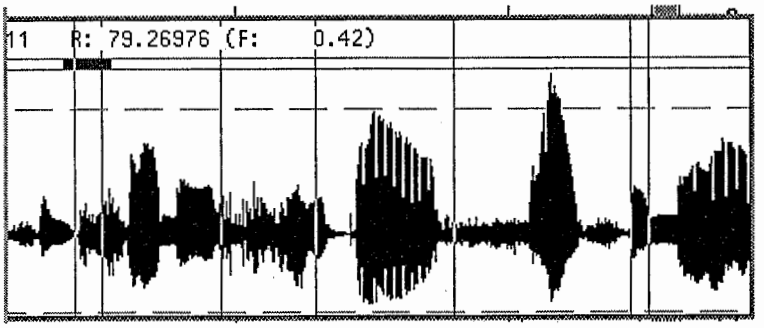
Different aspects of speech are transcribed for various purposes, the most universal levels of labelling are probably at the word-level (i.e., to yield a transcription of the words spoken), and at the phone-level (the fundamental speech sounds) to describe the utterance produced by the speaker.

Speech aligning is an essential precursor to tone and accent labelling under the ToBI system [1] (ToBI stands for Tones and Break indices), which is used for transcribing intonation patterns and other aspects of the prosody of English utterances. The ToBI system was devised to provide a standard for the production of a common set of prosodic elements in order to be able to share prosodically transcribed databases across research sites in the pursuit of diverse research purposes and varied technological goals. In our case, at ATR, the databases are subsequently used as raw material for synthesis units and for the study of prosodic aspects of human interaction. In order to be used for this purpose, the data needs to be sufficiently labelled with information describing the communicative nature of the speech. This is why prosodic labelling of the data is a necessary part of the process. In order to analyse and label a speech file, it must first be aligned accurately with the text words and phones.

3 Tools: the *Aligner* [2] and *waves+*

The speech signals are represented as plots of signal amplitude versus time. The appearance of the plot will depend on the type of speech or other sound contained within it. Syllables can be distinguished as egg-shaped periodic portions of the waveform, and consonants can be distinguished by the type of profile they display. Noisy parts of signal appear as random clumps, silences and quieter parts appear as a smoother continuous line.





3.1 Working with waves+

In this section we will examine some of the options available when viewing speech waves on a workstation using the ESPS [2, 3] program waves+ (xwaves). Waves+ is started by typing *xwaves filename* at the command line, where filename is the name of the speech file. By convention, files produced for viewing under waves+ have the suffix *.d*, or sometimes *.sd*.

- **Viewing the speech wave**

  Along the top of the window there is a purple strip running the whole length of the speech wave. Clicking the mouse in this strip enables movement along the file.

- To move forward one window-length into the file, click the left mouse button anywhere in the strip.

- To go back one window-length, click the right mouse button anywhere in the strip.

- To move to the last portion of a file, click the centre mouse button at the very right-hand end of the strip.

- To move to the first window, click the centre mouse button in the very first point of the strip.

- By clicking the right mouse button in the waves signal display, various options can be selected e.g., :

  Play window contents

  Play entire file

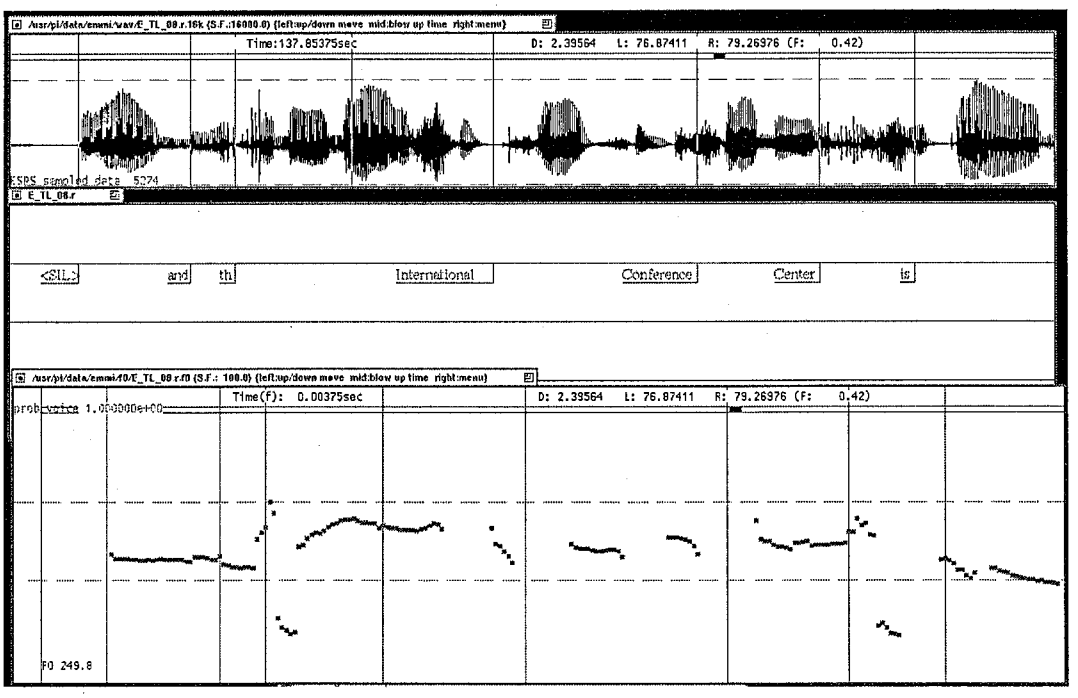
  Play between the marks.

  Here the "marks" are set by clicking the right mouse button to the left of the first word you want to listen to, a vertical dashed line will appear; with the mouse button still pressed, "drag" it to the end, i.e., to the right of the last word you want to listen to, and release it, another dashed line will appear at this point. Within the boundaries of these two dashed lines you will now have a "marked" part of speech. This can be listened to or zoomed into by selecting appropriate menu options. When the next part is marked in the same way, the previous "markings" will disappear. *Bracketing* a marked area means zooming in so that it fills the whole length of the window. It is quite common to *re-align* after bracketing to bring all associated windows into alignment at the same scale.

  Zoom in or out on a part of speech makes the size of the window over the speech waveform half or twice as large. This is useful for view more or less detail of the speech.

  N. B. Refer to the *xwaves* manual for a more in-depth explanation if necessary. It can be accessed by pressing the appropriate button in the main xwaves control panel. Alternatively, it can be called from the UNIX command line with the *eman* command.

  The *Aligner* automatically only displays 5 seconds of speech at any one time. This is because the detail would be too confused to analyse if the speech signal in one window was any larger.



4 Procedure

In this section we detail the most common procedures when labelling a speech file:

1. Aligning words in a speech utterance
2. Aligning and segmenting a speech utterance
3. Rejoining the aligned segment files as a new aligned complete file

4.1 Aligning words in a speech utterance

After digitisation, the speech file is viewed in xwaves. Since the Aligner is not robust to spontaneous speech, we first transcribe the words in the speech file and approximately time-align them to the speech wave. To do this, open *waves* and the speech file.

```
xwaves /DB/PI/data/speech-filename/wav/speech-filename.wav
```

```
(xwaves /DB/PI/data/sally/wav/sally01.wav)
```

Next open the label window. This can be done from the button in the control panel or from a shell script. The latter is more common.

- The words can be typed directly into the label text window. This involves a certain amount of guesswork, from viewing the speech wave pattern and listening to fragmented sections of the speech. The words do not have to be aligned accurately at this stage as it is done by running the automatic Aligner.
- Open the *Labeler* from the *xwaves* icon and change the *Label file* to a new filename for the word aligned file. The speech wave and an empty orthographic tier label strip should now appear on the screen. Type the words straight into the orthographic tier. When the window is closed, all the labels will be saved in a file noting their time-alignment to the speech waveform.

Such phenomena as filled pauses e.g., "um", can be transcribe in any way but the norm is to follow the ATIS corpus conventions, which specify "er", "mm", "uh", and "um" as the allowable transcriptions for filled pauses. It is always useful to have a *miscellaneous tier* open when labelling, so that notes can be jotted down and time-aligned to the speech being labelled.

4.2 Aligning and segmenting files from a speech utterance

Many words have more than one syllable, and therefore can be possibly uttered as stressed in more than one place. For this reason, not only is it important for the whole word to be accurately aligned with the uttered word with respect to the beginning and end boundaries, but also for the individual syllables in each utterance to be accurately distinguished.

1. Checking the alignment :

   The Aligner automatically selects the best phone sequence (word pronunciations) and time alignments for the speech utterance. However, the Aligner should not be expected to work well on recordings of noisy or distorted speech, hence the need to check and alter files after the automatic alignment. This is particularly important when aligning spontaneous speech, where the speaker naturally falters, stutters and pauses more irregularly than a non-spontaneous text read passage.

   Instead of attempting to align the whole speech file (often tens of minutes long) in one go, it is better to break it into smaller chunks. Small portions of speech are segmented from the complete file in order to align the syllables more accurately and avoid dealing with many windows of data at the same time. The number of segments depends on the size of the file. If the original complete file is small, segmenting may not be necessary at all. It is only required because the automatic routines sometimes get out-of-sync if they encounter too many non-speech noise and out-of-dictionary sequences.

2. Bring up the speech utterance file

```
xwaves /DB/PI/data/speech filename/wav/speech filename.wav

(xwaves /DB/PI/data/sally/wav/sally01.wav)
```

3. Mark and save a segment in a file

   - Mark a small section of speech, about 10 seconds long and select *"Play between the marks"* from the menu to take note of the content of the speech.
   - Activate the *waves* icon and at " OUTPUT file", name the new segment file. Change the filename in the *Labeler* window to new filename in order to save the file to be made separately from the original speech file. If the new filename ends with a number (*e.g., fred0*) then all subsequent saves will increment this number to save subsequent portions of speech as (*e.g., fred1.d, fred2.d, fred3.d*, etc.).
     (snt1)
     NOTE : Segment files should be numbered consecutively under the same filename to enable the segments to be rejoined together later for ToBI labelling.
   - Select *"Save segment in file"* from the menu
   - Quit out of *waves*. (Open the *waves* icon in the top left of the screen and select "Quit")
   - Open the new file under the new filename, check it has the same content as intended.

```
xwaves  filename.d

(xwaves bruce1.d)
```

4. Make the text file

Make a text file in a shell window for the whole original speech labels which can then be segmented into sentence fragments for more precise observation.

```
awk '{print $3}' label filename  >  text.

 (awk '{print $3}' snt/snt1.xlb  >  text.1)
```

this leaves the time information but selects the words and copies them to the text file.

Open the new text file from emacs (or whatever your favourite editor is). The text will appear in vertical column format, one word per line. In order to make it easier to read and execute, delete any "stray" lines from the beginning until the first line of the text starts at the top of the window. These extra lines come from the headers on the files. They are not important.

In emacs, **Esc Q** brings the text to a horizontal 'essay' type format ready to be segmented into smaller text files for aligning.

Highlight a few sentences from the text in one chunk:

In emacs, use **Esc W** to mark the beginning and the end of the chunk wanted.

Name and save the new file as a text (.txt) file. The new text filename should be named under the same name as the corresponding segment of speech in order for the aligner to be able to bring up the speech file with it's corresponding text at the same time for aligning.

Ctrl X Ctrl F filename.txt

(Ctrl X Ctrl F *bruce1.txt*)

Ctrl Y brings up the segmented text file.

Check the contents of the new text, bearing in mind any uncommon, slang or incomplete stuttered words which a standard dictionary would not recognise.

5. Make dictionary

If you haven't done so already, at this point create an executable dictionary directory in the output directory where unrecognised words can be entered. Every user of the aligner has his/her own dictionary of special words not found in the main dictionary.

mkdir /usr/local/align/dict/files/username.local.dic

Ctrl XD /usr/local/align/dict/files/fred.local.dic

Although there are automatic routines to help with this when the aligner is running, it is often faster (and more convenient) to enter any potentially unrecognisable words first by spelling, then by pronunciation, using the Aligner symbol set, (see Pronunciation Entries List in the Appendices), which is basically identical to the symbols for transcription adopted by the DARPA speech recognition community..

example:

vegetarianism, v eh jh ih t eh r iy ax n ih z m

coz, k aa z

NOTE: "words" in this context are not strictly proper words. The sounds often uttered in spontaneous speech, such as "errm", "ahh", can also be entered into the dictionary by pronunciation. Once entered into the dictionary, the same spelling of the sound should be maintained in subsequent word alignments to allow the dictionary to recognise it. This is best done by entering it in a menu.

6. Screening and automatic alignment

The batch mode program, *BAlign*, has two major phases of operation: data screening and automatic alignment. During data screening, the text file is checked to determine if the words it contains are in the Aligner dictionary. Each speech file is screened to be sure it exists and also the corresponding text file. Any problems at this stage are placed in a file called *BAproblems* in the output directory. This phase runs automatically and will complete the process by itself provided it doesn't come across any problems. When *BAlign* is finished, it prints notification on the terminal. The **.phones** and **.words** files can then be found in the output directory.

BAlign -S.d filename

The -S option is used to get the filename suffix (default .sd) to the more recent usage: .d; the text file suffix (default .txt) can be similarly reset by the -T. option.

(BAlign -S.d bruce1)

7. Align the data

Align is used to check the alignment result of a batch run of *BAlign*, or of somebody else's previous work.

Align -S.d filename

(Align -S.d bruce1)

8. Check the alignment

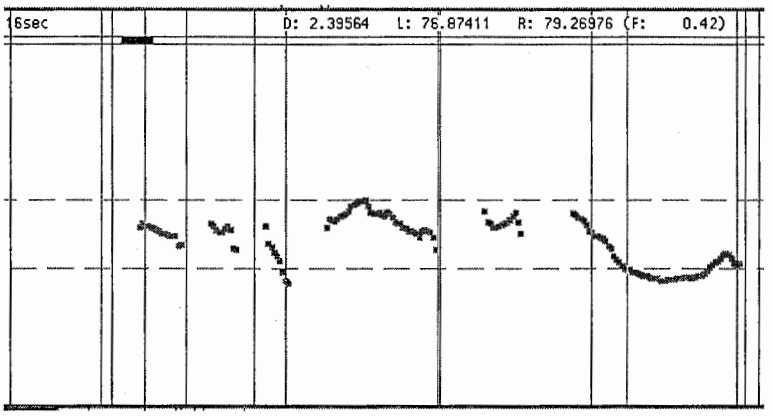
The new segment file can be listened to all at once, or in fragments using the drag-mouse technique described earlier to mark a grouping of several words, or in the label window by clicking the left mouse button on a particular word to listen to it by itself. This can be a very fast way of checking a file. Individual segments within a word can also be listened to individually in the same way from the segment-label window. Some of the segments will be mis-aligned; these can be corrected by listening to them and moving them accordingly. Select the MOVE option from the menu accessed by clicking on the right mouse button whilst in the segment/word display, then "picking up" the misplaced word using the centre mouse button and "dragging" it to the correct place. With practice, this can be a very fast operation.

4.3 Joining the aligned segment files as a new aligned complete file

Having segmented the speech file to make the aligning process easier to deal with, and completed the segmented files, the original label files need to be rejoined ready for *ToBI* labelling.

Three separate functions are needed :

1. **Phones** (the fundamental speech sounds)
2. **Words** (a transcription of the words spoken)
3. **f0 tone** (computer tracking of the voice pitch)



1. **Concatenate the phone files together**

```
cat filename?.phones > new-filename.phones

(cat bruce?.phones > spont1.phones)
```

(filename? includes all files of that name, e.g., bruce? includes bruce1, bruce2,.) from 0 to 9; if there are more than ten files to concatenate, use * instead of ? but beware of strange orderings under UNIX - sometimes 10 can come before 2!

The new filename will be the file containing all the aligned segment files in a continuous file. The speech of this new file is the same as the original speech utterance before it was segmented and aligned. This new file will be complete with speech (phones), words and f0 tone, and will be ready for ToBI Labelling.

2. **Concatenate the words together**

```
cat filename?.words > new-filename.words

(cat bruce?.words > spont1.words)
```

It is customary to put all label files in the same directory. They can be quite numerous, and should be separated by topic or group. Rather than copy the original speech files to the same directory (which can take up a lot of space) it is better to link them. Waves can then follow the links to read the original file.

ln -s /DB/PI/data/speech filename/wav/speech filename.wav new-name.d

(ln -s /DB/PI/data/sally/wav/sally01.wav spont1.d)

3. **"Get" the f0 (fundamental frequency, or 'pitch track')**

```
ls *f0

get_f0  new_filename.d  new_filename.f0

(get_f0  spont1.d  spont.f0)
```

N. B. The f0 tone may take a few minutes to "get".

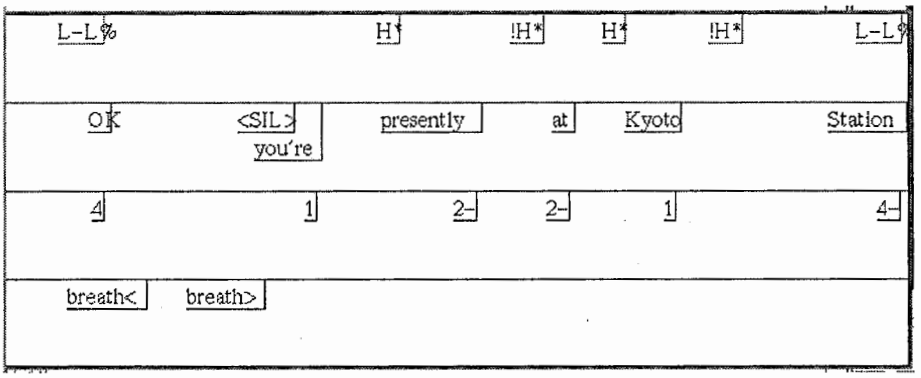
- Check in the editor window (e.g., *dired*) that the new file components are made, and have non-zero length.
- Open the new file:

(tobi spont1)

5 ENGLISH ToBI LABELLING GUIDELINES

In a completed file, there will be four levels of information:

- The Tone tier
- The Orthographic tier
- The Break Index tier
- The Miscellaneous tier



For more details see :

''Guidelines for ToBI Labelling''

(version 2. 0, February 1994)

by

Mary E. Beckman & Gayle M. Ayers [1]

Dired : /DB/PI/data/ToBI/labelling_guide_v2.ASCII<2>
(Also some visual examples are listed in the Appendix)

Or access it on the World Wide Web via :

http://www.itl.atr.co.jp/local_info/department/dept1/ToBI/main.html

5.1 THE TONE TIER

- **Boundary tones :**

  Also see Appendix for screen image guidelines and examples of labeling.

- L-L%

  For a full intonation phrase with an L phrase accent ending its final intermediate phrase and an L% boundary tone falling to a point low in the speaker's range.

- H-L%

  for an intonation phrase in which the H phrase accent of the final intermediate phrase upsteps the L% to a value in the middle of the speaker's range.

- L-H%

  for a full intonation phrase with an L phrase accent closing the last intermediate phrase, followed by an H boundary tone.

- H-H%

  for an intonation phrase with a final intermediate phrase ending in an H phrase accent and a subsequent H boundary tone.

  The boundary tones describe a definite boundary end to an intonation phrase and correspond to Break Index number 4

- **Phrasal tones :**

- %H

  High initial boundary tone, used only when a high pitch at the beginning of an utterance cannot be attributed to an H accent (H* or H+!H*) on the first or second syllable in the utterance

---

- **Phrase accents :**

- L- & H-

  The phrase accents describe intermediate phrase boundary and correspond to Break Index number 3

- **Pitch accents :**

  "Pitch accent tones are marked at every accented syllable. Lack of pitch accent assignment for a syllable will be interpreted as meaning that the syllable is NOT accented. "

- H*

  'peak accent' - an apparent tone target on the accented syllable which is in the upper part of the speaker's pitch range for the phrase.

- L+H*

  'rising peak accent' – a high peak target on the accented syllable which is immediately preceded by relatively sharp rise from a valley in the lowest part of the speaker's pitch range.

- L*

  'low accent' - an apparent tone target on the accented syllable which is in the lowest part of the speaker's pitch range.

- !H*

  preceding the downstepped pitch accent peak or downstepped H phrase accent. This diacritic is NEVER applied to the first H tone in a phrase.

- H+!H*

  a clear step down onto the accented syllable from a high pitch which itself cannot be accounted for by an H phrasal tone ending the preceding phrase or by a preceding H pitch accent in the same phrase; should only be used when the preceding material is clearly high-pitched and unaccented. (Otherwise the accent is a simple !H*.)

- L*+H

  'scooped accent' – a low tone target on the accented syllable which is immediately followed by relatively sharp rise to a peak in the upper part of the speaker's pitch range.

- HiF0

  To estimate a phrase's pitch range, mark a point within the pitch accent in the phrase which includes an 'H' tone and which contains the f0 maximum for the phrase.

5.2 THE BREAK INDEX TIER

Break Indices decribe the nature of the spacing in between words.

0, -1, 1p, 1, -2, 2, 2p, 3-, 3, 3p, -4, 4

- 0 : When two words are uttered so closely together that they have no space inbetween e. g, "got to" becomes "gotta".
- 1 : The most common index, describing a "normal" space in between words.
- 2 : a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; i. e. a well-formed tune continues across the juncture

  OR

  a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary.
- 3 : intermediate intonation phrase boundary; i. e. marked by a single phrase tone affecting the region from the last pitch accent to the boundary
- 4 : full intonation phrase boundary; i. e. marked by a final boundary tone after the last phrase tone.
- **Uncertainty :**

  Break indices -1, -2, etc describe uncertainty, e. g, -1 relates to uncertainty between 0 and 1; -2 relates to between 1 and 2 and so on.
- **The 'p' diacritic :**

  This is used for disfluency in timing, when a word is spoken hesitantly, as if searching for the next word, or when there is a prolongation of a word but no phrase accent.
- 1p : an abrupt cutoff before an actual repair or as if stopping to permit a repair or restart of some kind.
- 2p : a hesitation pause or prolongation of segmental material where there is no phrase accent perceived in the intonation contour
- 3p : a hesitation pause or a pause-like prolongation where there is a phrase accent in the tone tier.

5.3 THE MISCELLANEOUS TIER

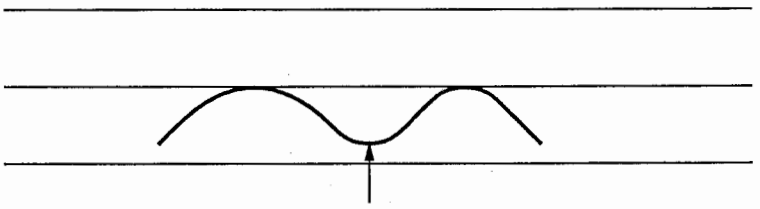
The miscellaneous tier will be used for any comments or markings (e. g, silence, audible breaths, laughter, disfluencies, and so on) desired by particular transcription groups.

6 JAPANESE ToBI LABELLING GUIDELINES

6.1 BREAK INDICES(BI)(English translation)

- BI1 : word (default specification) boundary ; includes particles
- BI2 : Accentual phrase (AP) boundary
- Break Index 2(i)

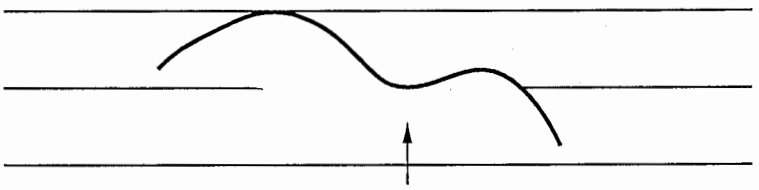
  In the case of two nuclear accents occurring one after the other and are of equal height:



BREAK INDEX 2

- Break Index 2(ii)

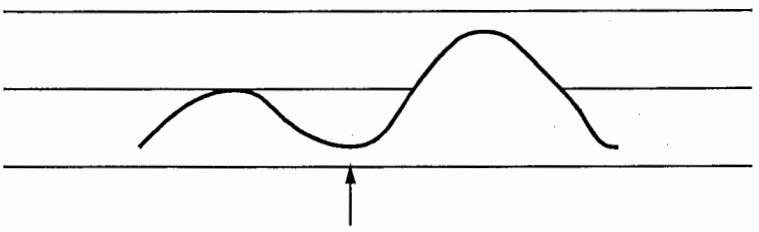
  In the case of two nuclear accents occurring one after the other, where the second accent is lower than the first :



BREAK INDEX 2

- BI3 : Intermediate phrase (IP) boundary

  In the case of two nuclear accents occurring one after the other, where the second accent is higher than the first :



BREAK INDEX 3

- BI4 : Intonation phrase (IP) boundary

  The end of the accentual phrase with a pause after it, or the end of the sentence

6.2 TONES (English translation)

- H*+L

  Marks lexical accent

  Place at the beginning of the pitch fall.

- H-

  Marks an unaccented accentual phrase or the "phrasal high" in an accented phrase, if this can be distinguished from the lexical accent in the F0 pattern.

  Place at the end of the initial pitch rise.

- W%L

  Post-pausal initial boundary tone

  Place just before the beginning of the accentual phrase

  Use in the case of either of the following :

  (1) there is a syllabic nasal sound on the second mora

  (2) the first accentual phrase begins with a long syllable

  (3) the first mora is accented

- %L

  Post-pausal initial boundary tone

  Place just before the beginning of the accentual phrase

  Use in the case of either of the following :

  (1) unaccented accentual phrases

  (2) the second mora or the remaining mora is/are accented

- WL%

  Weak final boundary tone of accentual phrase

  Place at the same interval as break index 2 or 3

  Usage : refer to (1), (2), (3) of W%L above

- L%

  Final boundary tone of accentual phrase

  Place just at the break index 2, 3 or 4

  Break index 2 or 3 : refer to (1) & (2) mentioned in %L above

  Break index 4 : use in the case of no final rise

- H%

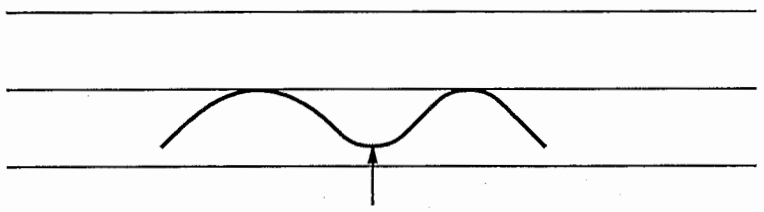
  Final boundary tone of intonational phrase

  Place at the same interval as break index 4

  Use in the case of a final rise

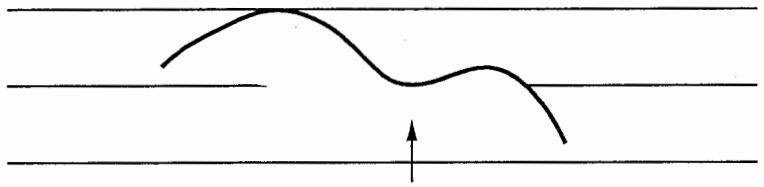
6.3 BREAK INDEX (Japanese)

- BI 1 ：単語、助詞の境界に置く。
- BI 2 ：アクセント句の境界に置く。
- Break Index 2(i)
  境界前後のアクセント核の F0 が、ほぼ同じ場合：



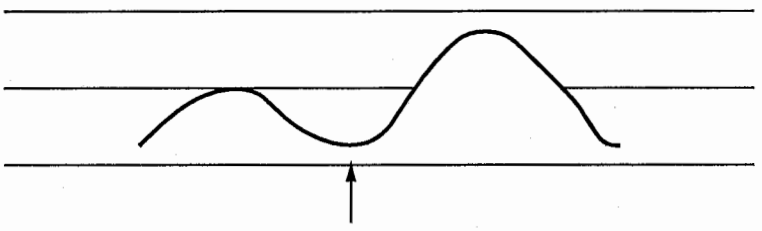
BREAK INDEX 2

- Break Index 2(ii)
  後ろのアクセント核の F0 方が低い場合：



BREAK INDEX 2

- BI 3 ：アクセント句の境界に置く。
  - 境界後ろのアクセント核の F0 が、前のアクセント核の F0 より、高い場合：



BREAK INDEX 3

- BI 4 ：長いポーズのある文節境界及び、文末に置く。

6.4 TONES (Japanese)

- H*+L

  ピッチの下降始めに置く。

  「起伏型アクセント句」に用いる。

- H-

  ピッチの上昇終りに置く。

  「1型、2型以外のアクセント句」に用いる。

- W%L

  文頭または、BREAK INDEX 4 の後の文節頭（「##」上）に置く。

  次にくるアクセント句が (1)(2)(3) のいずれかの場合に用いる。

  (1) 2 mora 目が「撥音」である

  (2) 1 または 2 mora 目が「長母音」を構成する

  (3) アクセント型が 1 型である

- %L

  文頭または、BREAK INDEX 4 の後の文節頭（「##」上）に置く。

  次にくるアクセント句が (1)(2) のいずれかの場合に用いる。

  (1) 平板型である

  (2) 1 型以外の「起伏型アクセント」である

- WL%

  アクセント句の終り（BREAK INDEX 2 or 3 の上）に置く。

  使用条件は上記 W%L の (1)(2)(3) と同じ。

- L%

  アクセント句の終り（BREAK INDEX 2, 3 or 4 の上）に置く。

  （BREAK INDEX 2 or 3 の場合）

  使用条件は上記 %L の (1)(2) と同じ。

  （BREAK INDEX 4 の場合）

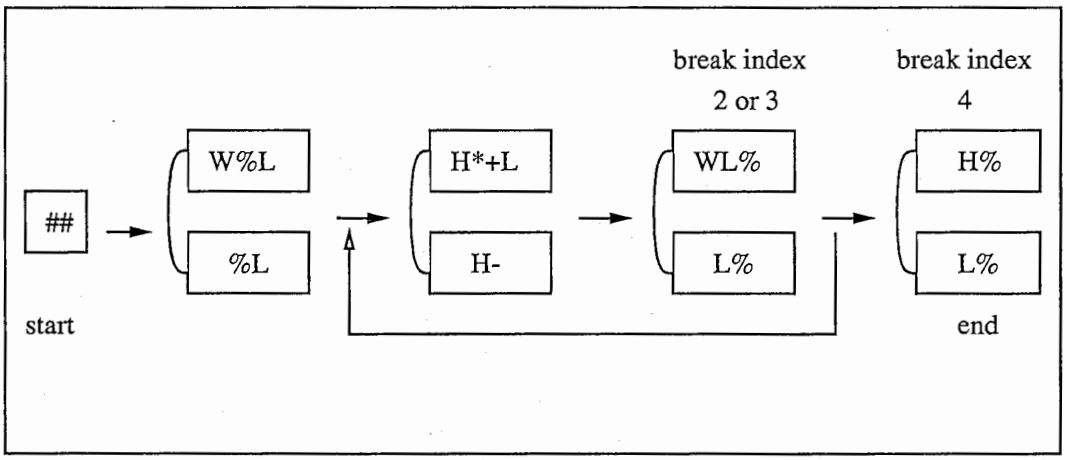
  BREAK INDEX 直前が上がり調子でない場合に用いる。

- H%

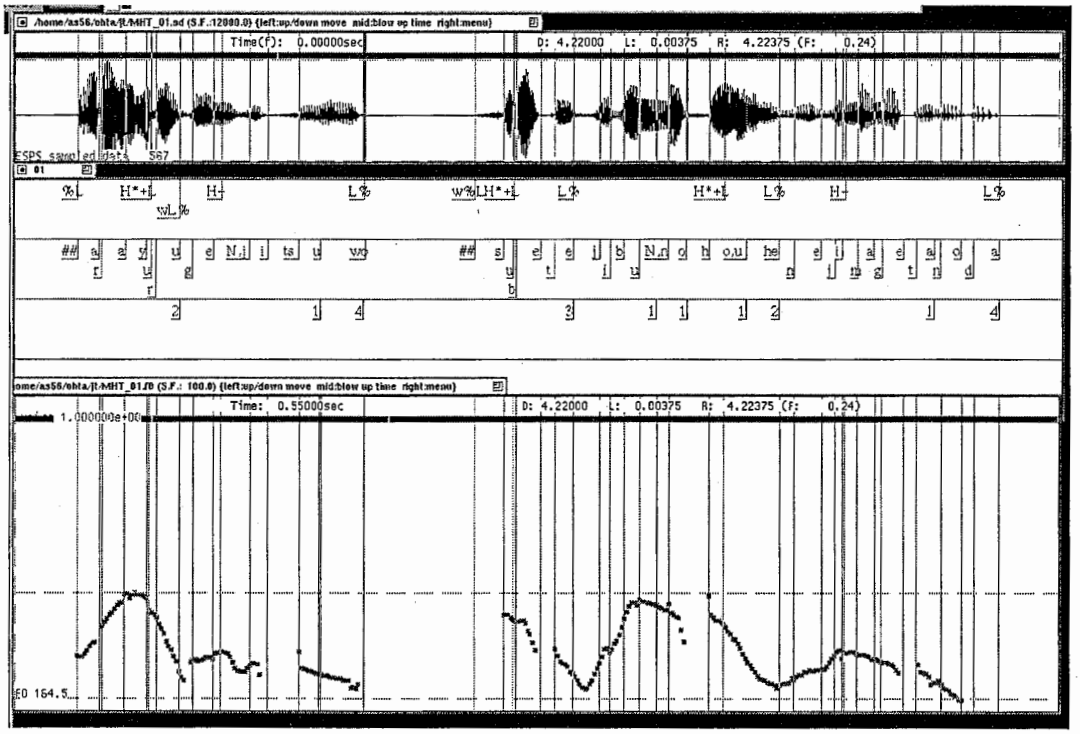
  イントネーション句の終り（BREAK INDEX 4 の上）に置く。

  BREAK INDEX 直前が上がり調子の場合に用いる。

- Flow chart of tone labels



- Sample labelling



7 KOREAN LABELLING GUIDELINES

韓国語の文字ハングルは、表音文字である。ハングル文字は３つの要素（初声：子音、中声：音節核一母音、終声：子音）の組合せで構成される。つまり、

Ci + V + (Cf)：ハングル一文字の音節構造

Ci ： 初声（子音—文字１９ 、音素１８個）
V ： 音節核（単母音９個、半母音（w、 y）＋単母音１１個、複母音１個）
Cf ： 終声（子音—文字２７、 音素７個）

ちなみに、（ ）は、（ ）内のそれがなくても、文字として成立する場合もあることを表している。また、初声（子音）には、 V と VC 型を表すための /ㅇ/ を含めて、１９種類の文字がある。 /ㅇ/ とは、いわゆる黙音で、表記はされるが発音されない文字である。さらに、終声（子音）については２７種類の文字があるが、これは音韻としては、７種類の音素に帰着する。

各々の音節要素に関して、表１）に韓国語の音素表を挙げる。ハングル表記、ラベリングにおいての記号、調音様式と調音位置が示されている。表における音素発声について、以下に、説明を付加する。

初声の音素の発声は、

/ㄱ, ㅂ, ㄷ(g,b,d)/ : voiceless unaspired lenis plosive
: 有声音の間では有声化
/ㅈ(z)/ : voiceless unaspired lenis affricative
: 有声音の間では有声化
/ㅅ(s)/ : lenis voiceless alveolar fricative
/ㄲ, ㄸ,ㅃ, ㅆ, ㅉ(K,T,P,S,Z)/ : voiceless unaspirated fortis sound
/ㅋ, ㅌ, ㅍ, ㅊ(k,t,p,c)/ : strongly aspirated voiceless sound

終声の音素の発声は、

/ㄱ(G)/ : voiceless velar stop without plosion
/ㄷ(D)/ : voiceless alveolar stop without plosion
/ㅂ(B)/ : voiceless bilabial stop without plosion
/ㄴ, ㅁ, ㅇ, ㄹ(N,M,Q,R)/ : sonorant 文献１）

また、韓国語では、音素が音韻環境により他の音素に変わる音韻変動や音韻帰着があるため、正書法文表記と音素表記とが異なる。以下に、例文をそれぞれの表記法で示す。

：##表記例##

- 正書法文表記 (Orthographic version):

바람과 햇님이 서로 힘이 더 세다고 다투고 있을 때, 한 나그네가 따뜻한 외투를 입고 걸어 왔습니다.

- 音韻変動後の音素表記 (phoneme version) :

바람과 핸니미 서로 히미 더 세다고 다투고 이쓸 때, 한 나그네가 따뜨탄 외투를 입꼬 거러 와씀니다.

- ラベリングでの対応記号で表記 (labeling symbol version):

baraMgwa hANnimi sEro himi sedago datugo iSIR TA, haN nagInega TaTItaN weturIR iBKo gErEwaSIMnida.

表 1：韓国語の音素表

音節要素	音素 ハングル	ラベリング 対応記号	IPA 記号	調音様式	調音位置
初声 (子音)	ㄱ	g	g	plosive	velars
	ㄴ	n	n	nasal	alveolars
	ㄷ	d	d	plosive	alveolars
	ㄹ	r	l	liquid	alveolars
	ㅁ	m	m	nasal	labials
	ㅂ	b	b	plosive	labials
	ㅅ	s	s	fricative	alveolars
	ㅈ	z	ʤ	affricate	palatals
	ㅊ	c	c^h	affricate	palatals
	ㅋ	k	k^h	plosive	velars
	ㅌ	t	t^h	plosive	alveolars
	ㅍ	p	p^h	plosive	labials
	ㅎ	h	h	fricative	glottal
	ㄱㄱ	K	k	plosive	velars
	ㄷㄷ	T	t	plosive	alveolars
	ㅂㅂ	P	p	plosive	labials
	ㅅㅅ	S	s	fricative	alveolars
	ㅈㅈ	Z	c	affricate	palatals
中声 (音節核)	ㅏ	a	a	low	central unrnd
	ㅓ	E	ʌ	mid	central unrnd
	ㅗ	o	o	higher mid	back round
	ㅜ	u	u	high	back round
	ㅡ	I	ɯ	high	central unrnd
	ㅣ	i	i	high	front unrnd
	ㅏㅣ	A	ɛ	lower mid	front unrnd
	ㅓㅣ	e	e	higher mid	front unrnd
	ㅗㅣ	O(w+e)	ɸ	higher mid	front round
	ㅑ	y+a	ja		
	ㅕ	y+E	jʌ		
	ㅛ	y+o	jo		
	ㅠ	y+u	ju		
	ㅣㅓㅣ	y+e	je		
	ㅣㅏㅣ	y+A	jɛ		
	ㅗㅏ	w+a	wa		
	ㅜㅓ	w+E	wʌ		
	ㅗㅏㅣ	w+A	wɛ		
	ㅗㅓㅣ	w+e	we		
	ㅜㅣ	w+i	wi		
	ㅡㅣ	U	ɯi		

表 1：韓国語の音素表（継き）

音節要素	音素 ハングル	ラベリング 対応記号	IPA 記号	調音様式	調音位置
終声 (子音)	ㄱ	G	k̚	plosive	velars
	ㄴ	N	n	nasal	alveolars
	ㄷ	D	t̚	plosive	alveolars
	ㄹ	R	l	liquid	alveolars
	ㅁ	M	m	nasal	labials
	ㅂ	B	p̚	plosive	labials
	ㅇ	Q	ŋ	nasal	velars

8 Pronunciation entries

The aligner offers several ways to update the dictionary or correct a text entry. You can correct typographic errors in the transcription by editing the text in the "Word-level transcription" window. Press the "Fix Transcription" button to register the change.

You can add a new pronunciation by entering its space-separated symbols on the "Pronunciation" line. Press the "Update Dictionary" button when you have entered the pronunciation. You may browse the "Current Dictionary Contents" for similar pronunciations and double click to insert one that is close. Syllable boundaries, indicated by space-separated "*" are optional.

New words with regular inflection are best entered as baseforms, rather than inflected forms. Thus, if the missing word is "helping", enter "help" and its pronunciation, rather than "helping". To reduce an inflected unknown to its base form, change it on the "Unknown word" line.

Symbols listed in the first column of the following table may be used in the "Pronunciation" entry. Note that these are essentially identical to the DARPA symbols, except for the absence of "er", "nx" and "dx". Approximate TIMIT equivalents are shown only for reference.

Aligner	TIMIT	DARPA	EXAMPLE
aa	aa	aa	cot
ae	ae	ae	bat
ay	ay	ay	bite
aw	aw	aw	now
b	bcl, b	b	bob
ao	ao	ao	bought
ch	ch	ch	church
oy	oy	oy	boy
d	dcl, d	d	dad
dh	dh	dh	they
eh	eh	eh	bet
ey	ey	ey	bait
er	er	er	bird
ah, r	er	er	bird
ax, r	er	er	bird
f	f	f	fief
g	gcl, g	g	gag
hh	hh	hh	hay (unvoiced h)
hh	hv	hh	hay (voiced h)
ax	ix	ix	roses
ih	ih	ih	bit
iy	iy	iy	beat
jh	jh	jh	judge
k	kcl, k	k	kick
l	l	l	led
el	el	el	bottle (syllabic l)

Pronunciation entries continued

Aligner	TIMIT	DARPA	EXAMPLE
m	m	m	mom
eh, m	em	m	'em (syllabic m)
n	n	n	nun
n, t	nx	nx	center (nasal flap)
ng	ng	ng	sing
ng	eng	ng	sing
en	en	en	button (syllabic n)
ow	ow	ow	boat
p	pcl, p	p	pop
r	r	r	red
axr	axr	axr	dinner
ax, r	ax	r axr	dinner
ah, r	axr	axr	dinner
s	s	s	sister
sh	sh	sh	shoe
dx	dx	dx	butter (alveolar flap)
t	tcl, t	t	tot
th	th	th	thief
uh	uh	uh	book
ah	ah	ah	butt
ax	ax	ax	the (schwa)
ax	ax-h	ax	the (unvoiced schwa)
uw	uw	uw	boot
uw	ux	uw	beauty
v	v	v	verve
w	w	w	wet
y	j	y	yet
y	y	y	yet
z	z	z	zoo
zh	zh	zh	measure
SIL	sil	sil	silence
SIL	h#	sil	silence
SIL	pau	sil	silence

9 File/Directory listings

9.1 SPONT

Spont is a monologue in which the speaker talks freely about her impressions of Japan.

```
Dired :  /DB/PI/data/SPONT
```

Files cited in this report (snt; bruce; spont), can be found in the SPONT directory

9.2 emmi

Emmi is a set of dialogues taken from a multi-media experiment. The content of the files is the conversations that took place between a travel agent and a client, the agent is advising the client on various ways to get to the International Conference Centre. There are two types of *emmi* data :-

```
TL    : Telephone Line

MM    : Multi-Modal

Dired : /usr/pi/data/emmi/ToBI-lbl/
```

There are 10 files for MM and 10 for TL, each type has two sides of the conversation. They were originally named after the channel in the stereo signal that each was recorded on:

- r = agents voice (right speaker). This half of the conversation is the most detailed and is the data used for analysis.
- l = customers voice (left speaker)

9.3 Communicative Act Labels

The same EMMI data was used as the base on which to build a set of Communicative Act Labels.

```
For files and further information, see :

Dired : /DB/PI/data/emmi/ift

Technical report no.:  TR-IT-0081

Title : ''A Bilingual Set of Communicative Act Labels for Spontaneous
        Dialogues''
```

Acknowledgements

The authors would like to thank Prof. Lee of Dong-duck Womens' University, Korea, and Sinhwa Kang for their help and advice about Korean labelling. We are also extremely grateful to Dr Yasuhiro Yamazaki and Dr Norio Higuchi for their kind support of this work.

References

[1] M. E. Beckman, G. M. Ayers *Guidelines for ToBI Labelling* - version 2. 0, February 1994.

[2] Colin Wightman, David Talkin *The Aligner: A system for automatic time alignment of English text and speech* Document version : 1. 3 of 3/11/94 Entropic Research Laboratory Inc. 1994.

[3] Entropic Research Laboratory Inc. AT & T Bell Laboratories *waves+* 1993 Version 3 . 1

Appendix 1

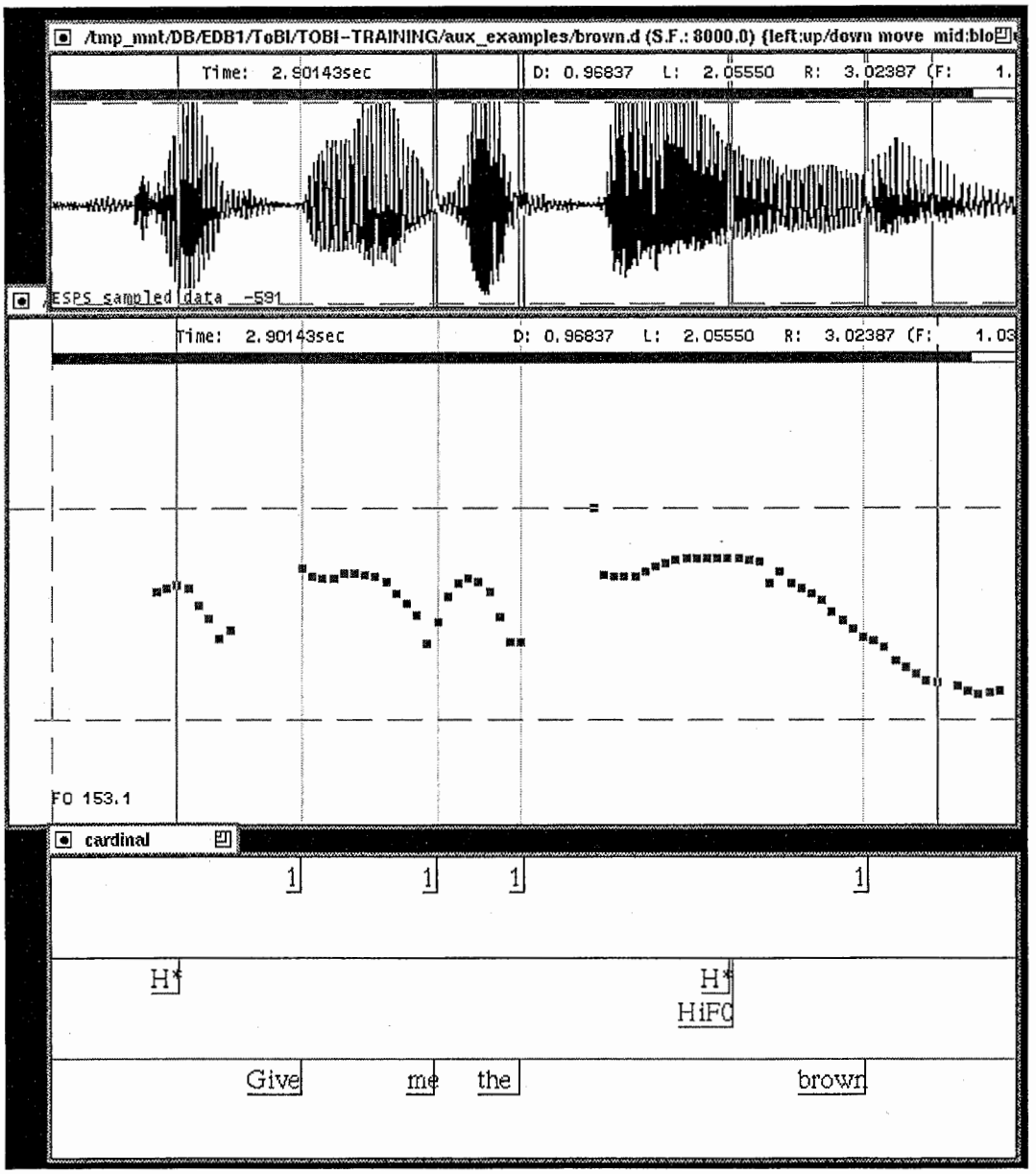
English ToBI labeling guidelines

More examples can also be accessed on the World Wide Web via :

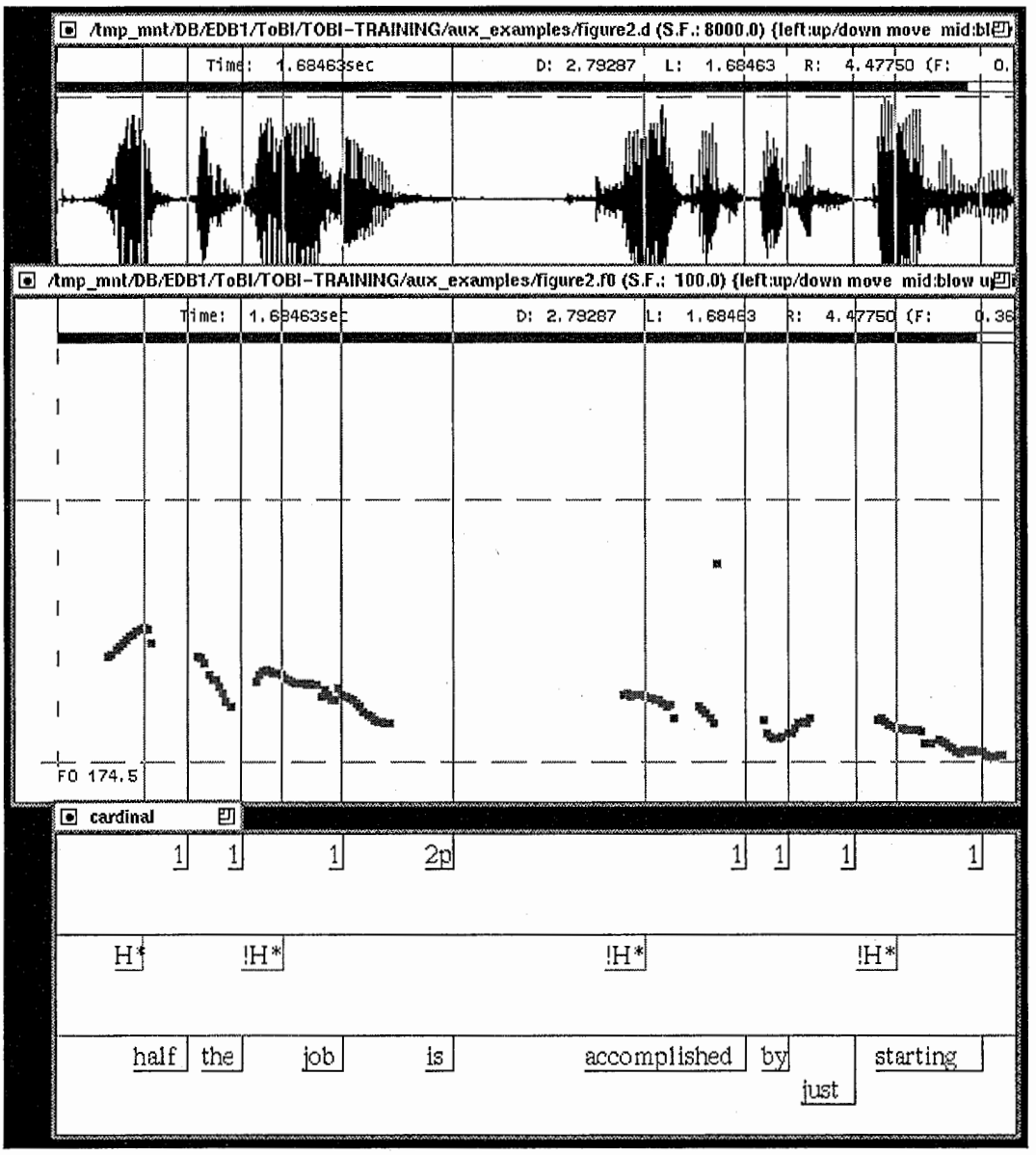
`http://www.itl.atr.co.jp/local_info/department/dept2/ToBI/main.html`

0.1 Pitch accents

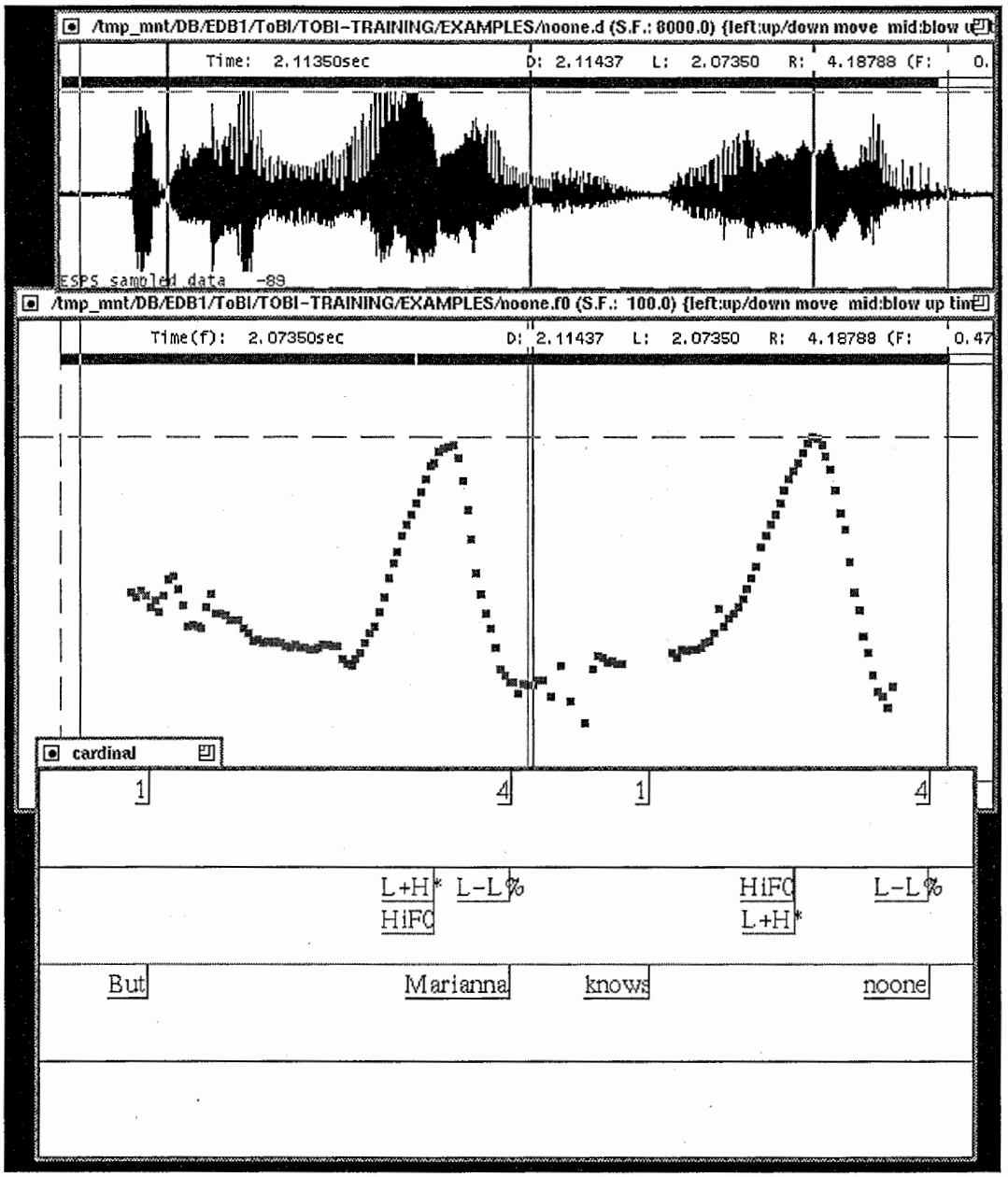
- H*



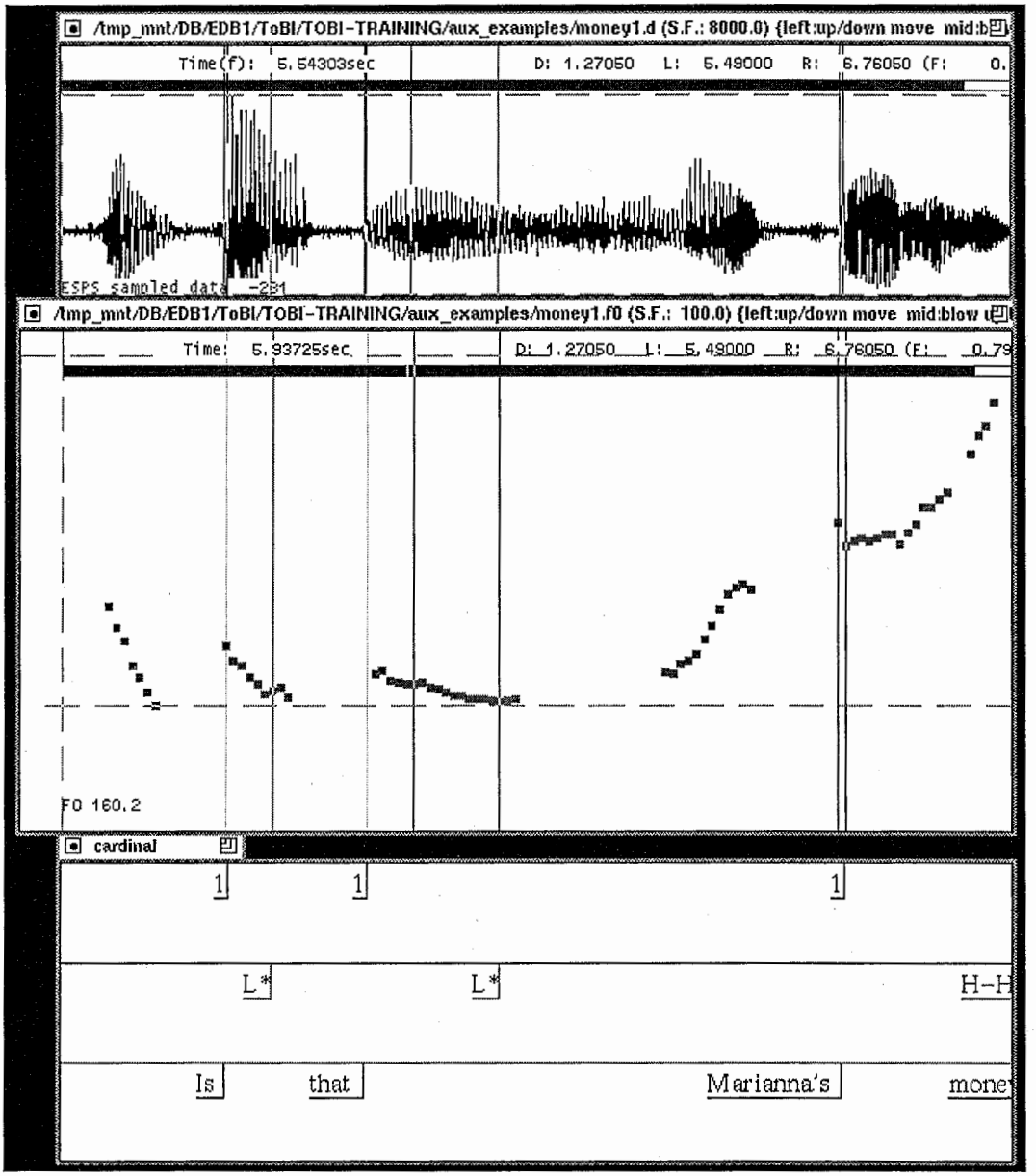
- !H*


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/figure2.d (S.F.: 8000.0) {left:up/down move mid:bl
Time: 1.68463sec
D: 2.79287
L: 1.68463
R: 4.47750 (F:
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/figure2.f0 (S.F.: 100.0) {left:up/down move mid:blow u
Time: 1.68463sec
D: 2.79287
L: 1.68463
R: 4.47750 (F:
F0 174.5
cardinal
1 1 1 2p 1 1 1 1
H* !H* !H* !H*
half the job is accomplished by just starting

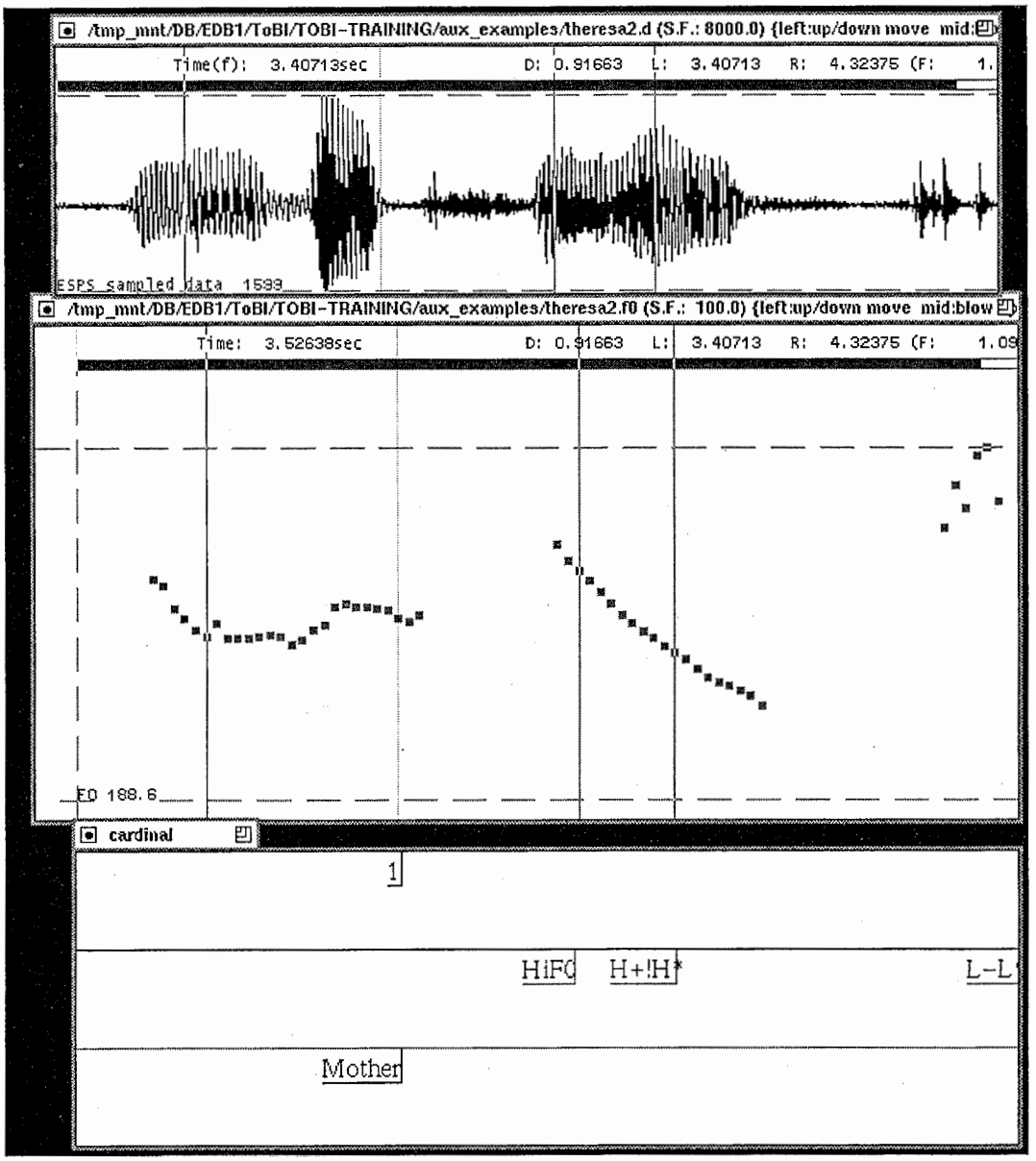
- L+H*


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/noone.d (S.F.: 8000.0) {left:up/down move mid:blow
Time: 2.11350sec
D: 2.11437 L: 2.07350 R: 4.18788 (F:
ESPS sampled data -89
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/noone.f0 (S.F.: 100.0) {left:up/down move mid:blow up tim
Time(f): 2.07350sec
D: 2.11437 L: 2.07350 R: 4.18788 (F: 0.47
cardinal
1
4
1
4
L+H*
HiFO
L-L%
HiFO
L+H*
L-L%
But
Marianna
knows
noone

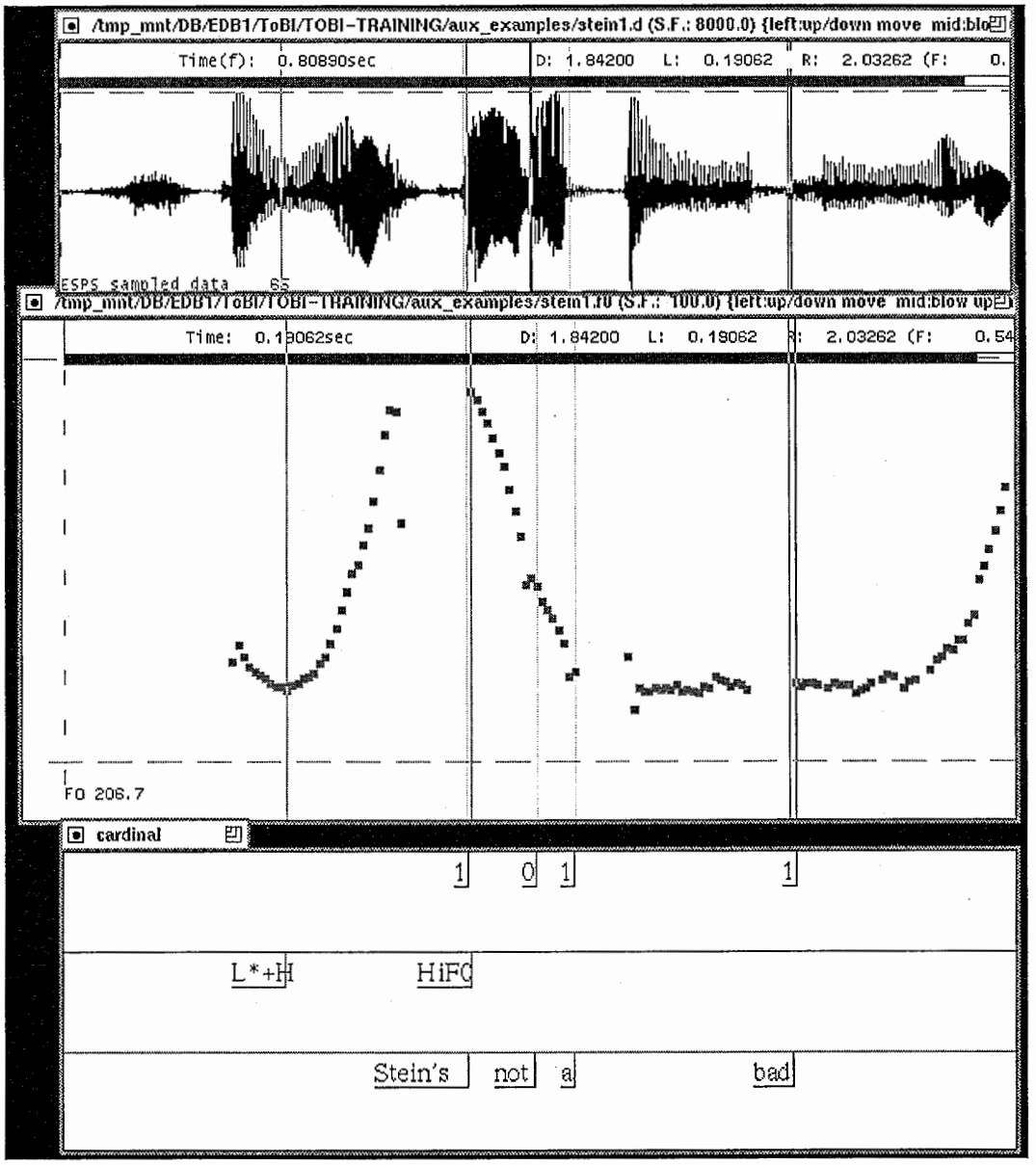
- L*


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/money1.d (S.F.: 8000.0) {left:up/down move mid:b
Time(f): 5.54303sec
D: 1.27050 L: 5.49000 R: 6.76050 (F:
ESPS sampled data -2B1
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/money1.f0 (S.F.: 100.0) {left:up/down move mid:blow
Time: 5.93725sec
D: 1.27050 L: 5.49000 R: 6.76050 (F:
F0 160.2
cardinal
1
1
1
L*
L*
H-H
Is
that
Marianna's
mone

- H+!H*

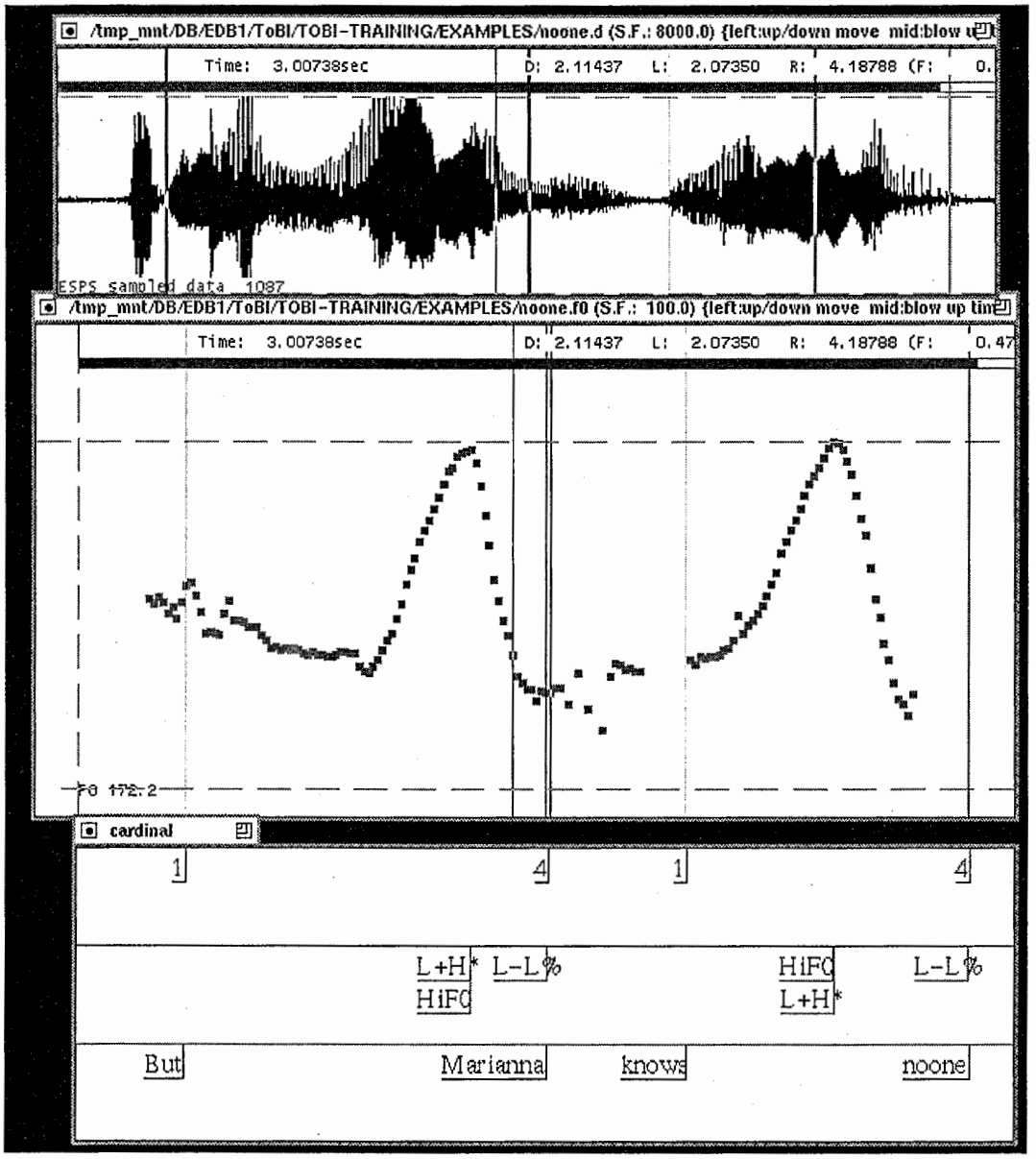

/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/theresa2.d (S.F.: 8000.0) {left:up/down move mid:
Time(f): 3.40713sec
D: 0.91663 L: 3.40713 R: 4.32375 (F: 1.
ESPS sampled data 1599
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/theresa2.f0 (S.F.: 100.0) {left:up/down move mid:blow
Time: 3.52638sec
D: 0.91663 L: 3.40713 R: 4.32375 (F: 1.09
ED 188.6
cardinal
1
HiF0
H+!H*
L-L
Mother

- L*+H

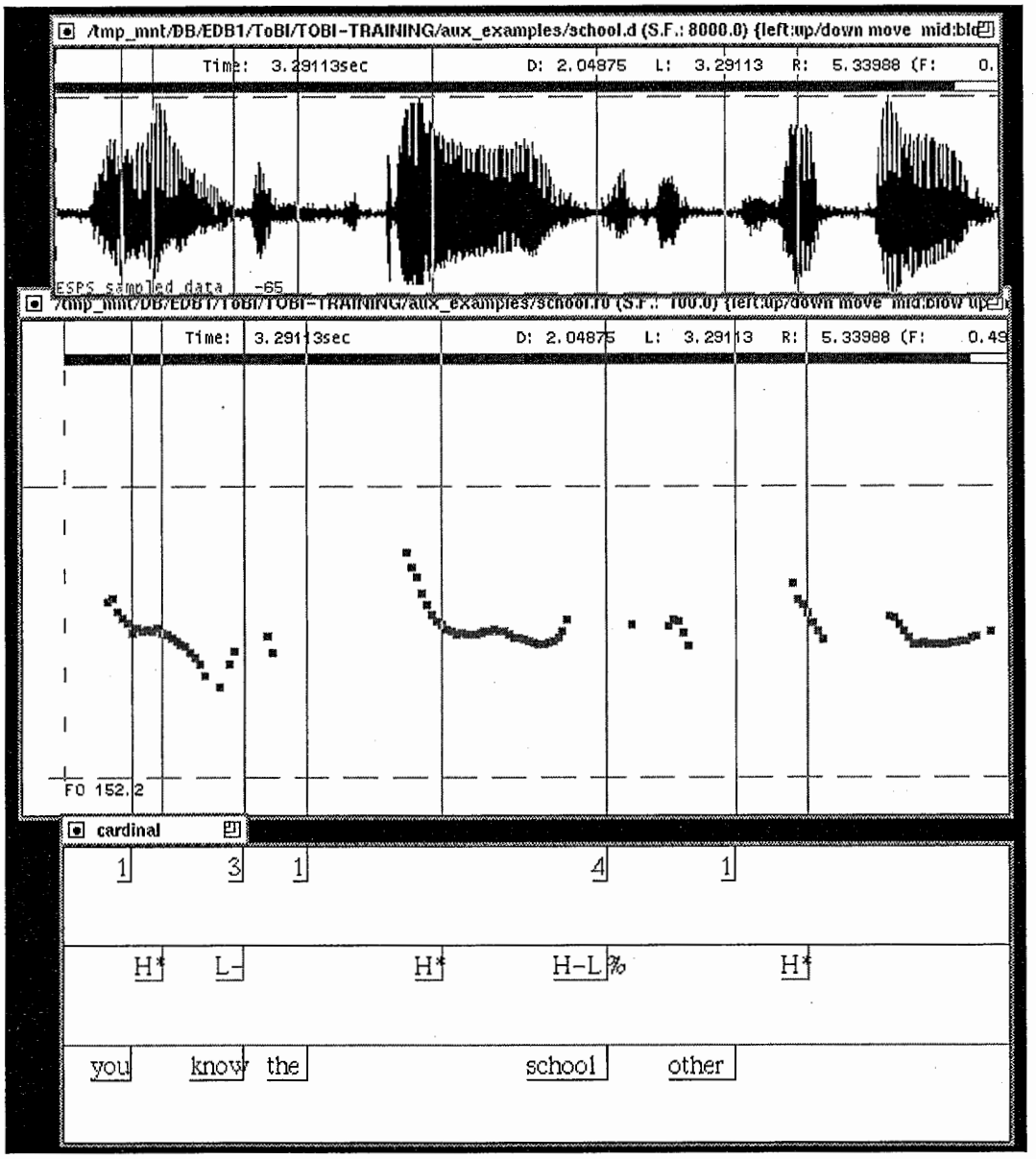

/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/stein1.d (S.F.: 8000.0) {left:up/down move mid:blo
Time(f): 0.80890sec
D: 1.84200 L: 0.19062 R: 2.03262 (F: 0.
ESPS sampled data 65
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/stein1.f0 (S.F.: 100.0) {left:up/down move mid:blow up
Time: 0.13062sec
D: 1.84200 L: 0.19062 R: 2.03262 (F: 0.54
F0 206.7
cardinal
1 0 1 1
L*+H HiF0
Stein's not a bad

Boundary tones

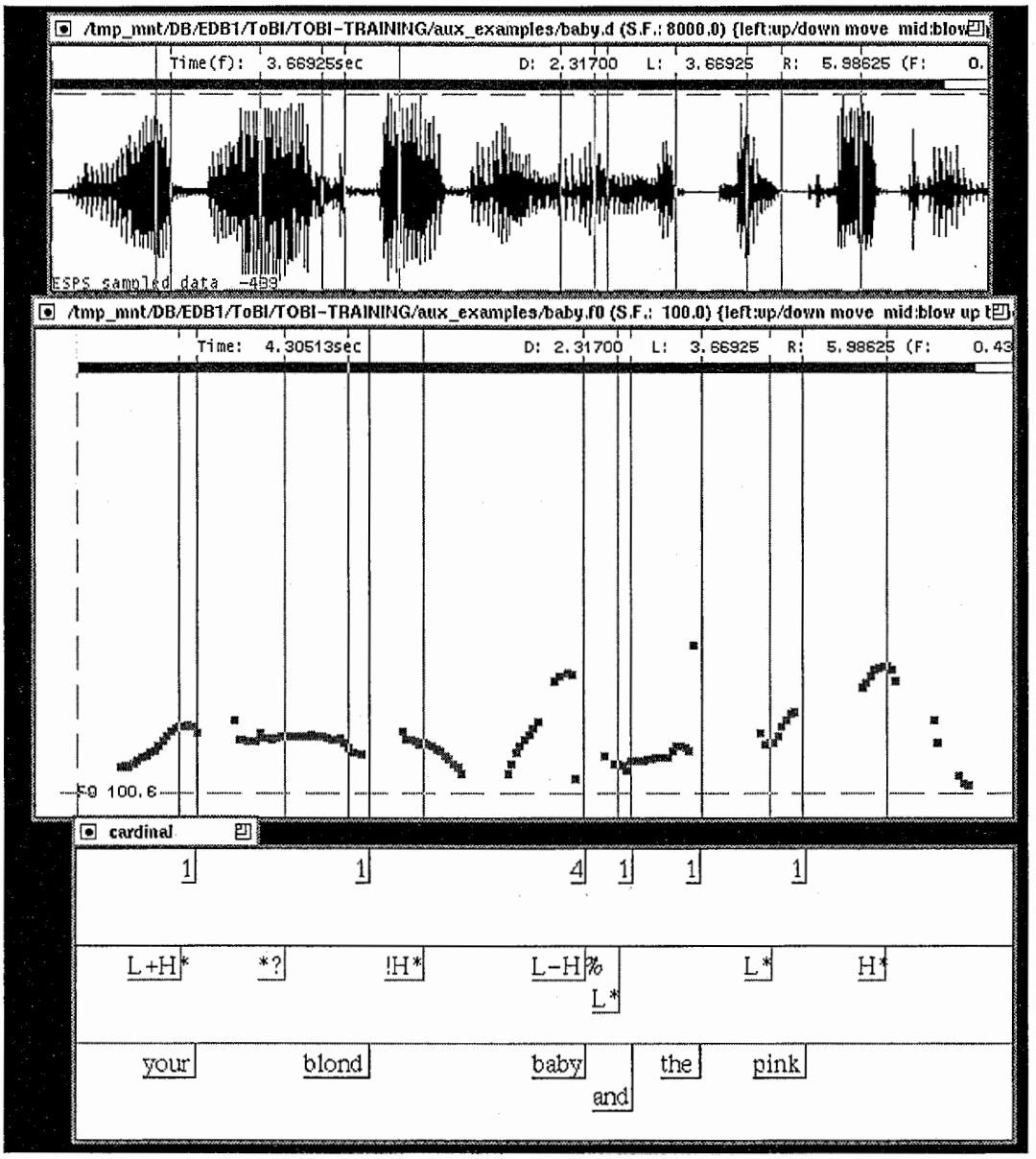
- L-L%



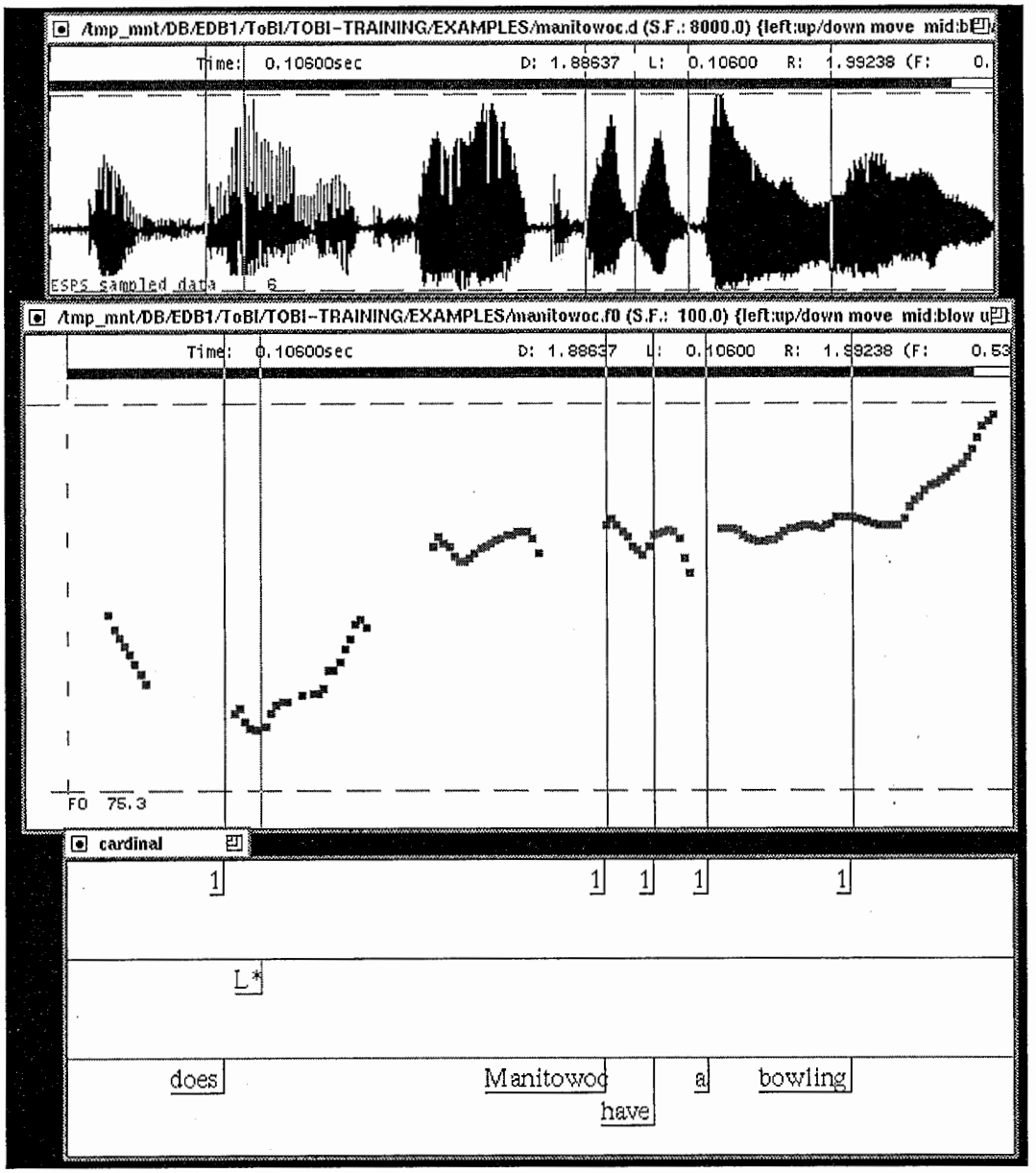
- H-L%


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/school.d (S.F.: 8000.0) {left:up/down move mid:bl
Time: 3.29113sec
D: 2.04875 L: 3.29113 R: 5.33988 (F: 0.
ESPS sampled data -65
Time: 3.29113sec
D: 2.04875 L: 3.29113 R: 5.33988 (F: 0.49
F0 152.2
cardinal
1
3
1
4
1
H*
L-
H*
H-L%
H*
you
know
the
school
other

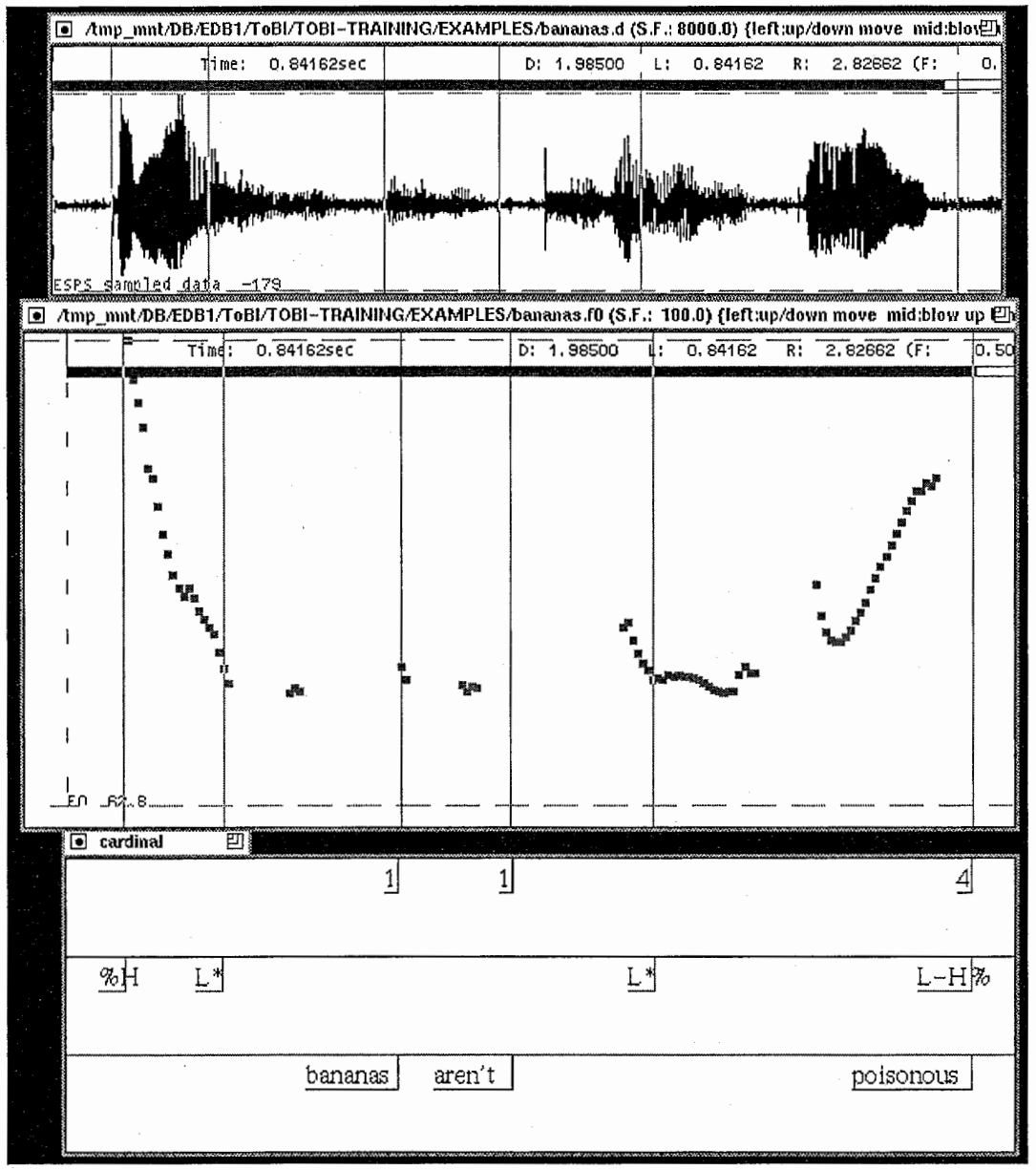
- L-H%


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/baby.d (S.F.: 8000.0) {left:up/down move mid:blow
Time(f): 3.66925sec
D: 2.31700 L: 3.66925 R: 5.98625 (F: 0.
ESPS sampled data -4B9
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/aux_examples/baby.f0 (S.F.: 100.0) {left:up/down move mid:blow up
Time: 4.30513sec
D: 2.31700 L: 3.66925 R: 5.98625 (F: 0.43
F0 100.6
cardinal
1 1 4 1 1 1
L+H* *? !H* L-H% L* H*
L*
your blond baby and the pink

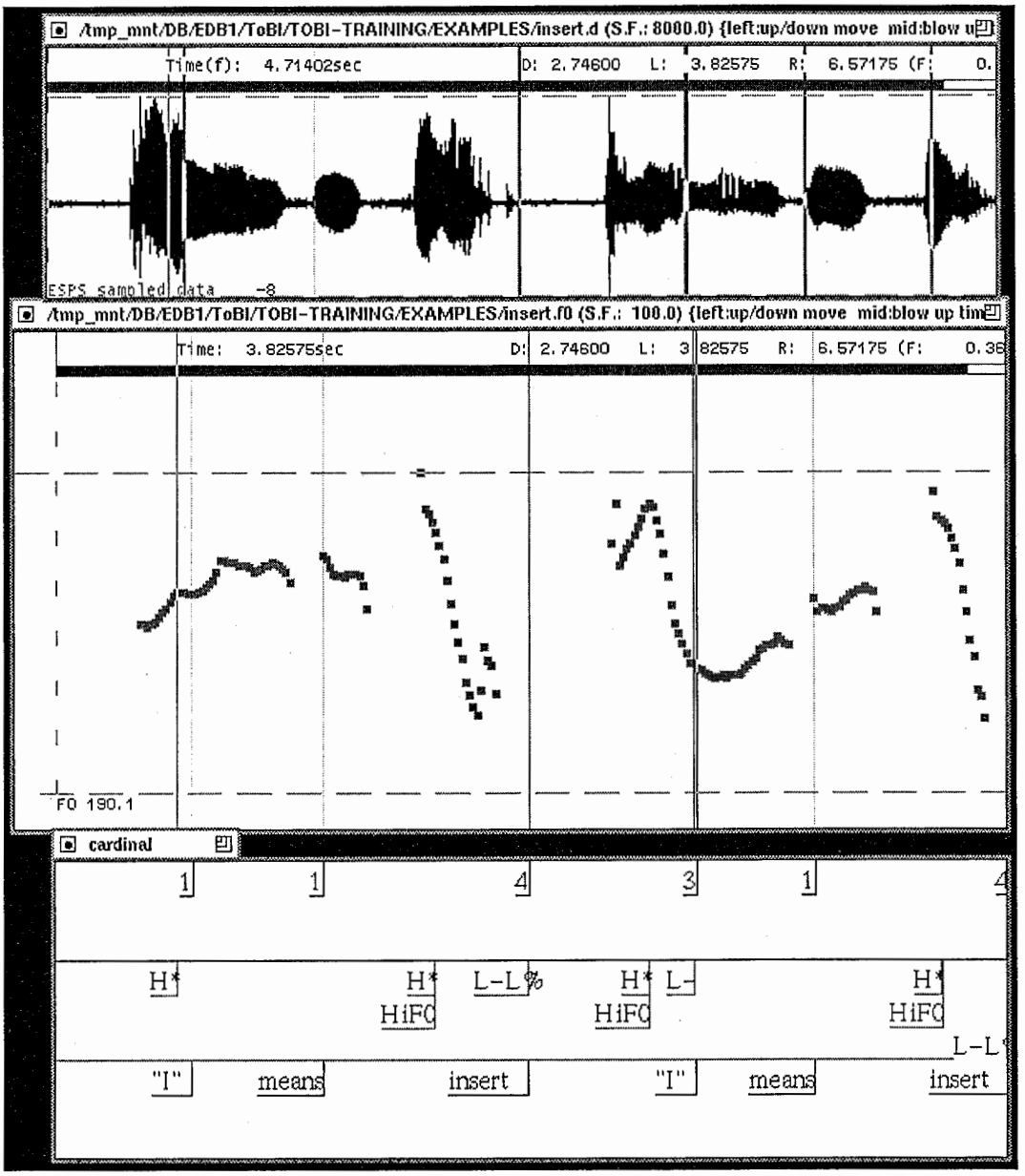
- H-H%


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/manitowoc.d (S.F.: 8000.0) {left:up/down move mid:b
Time: 0.10600sec D: 1.88637 L: 0.10600 R: 1.99238 (F: 0.
ESPS sampled data
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/manitowoc.f0 (S.F.: 100.0) {left:up/down move mid:blow u
Time: 0.10600sec D: 1.88637 L: 0.10600 R: 1.99238 (F: 0.53
F0 75.3
cardinal
1 1 1 1 1
L*
does Manitowoc a bowling
have

- %H


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/bananas.d (S.F.: 8000.0) {left:up/down move mid:blow
Time: 0.84162sec
D: 1.98500 L: 0.84162 R: 2.82662 (F:
ESPS sampled data -179
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/bananas.f0 (S.F.: 100.0) {left:up/down move mid:blow up
Time: 0.84162sec
D: 1.98500 L: 0.84162 R: 2.82662 (F:
0.50
cardinal
1
1
4
%H
L*
L*
L-H%
bananas
aren't
poisonous

- L-


/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/insert.d (S.F.: 8000.0) {left:up/down move mid:blow u
Time(f): 4.71402sec
D: 2.74600 L: 3.82575 R: 6.57175 (F:
ESPS sampled data -8
/tmp_mnt/DB/EDB1/ToBI/TOBI-TRAINING/EXAMPLES/insert.f0 (S.F.: 100.0) {left:up/down move mid:blow up tim
Time: 3.82575sec
D: 2.74600 L: 3.82575 R: 6.57175 (F:
F0 190.1
cardinal
1 1 4 3 1 4
H* H* L-L% H* L- H*
HiF0 HiF0 HiF0
L-L
"I" means insert "I" means insert

Appendix 2

English ToBI labeling examples

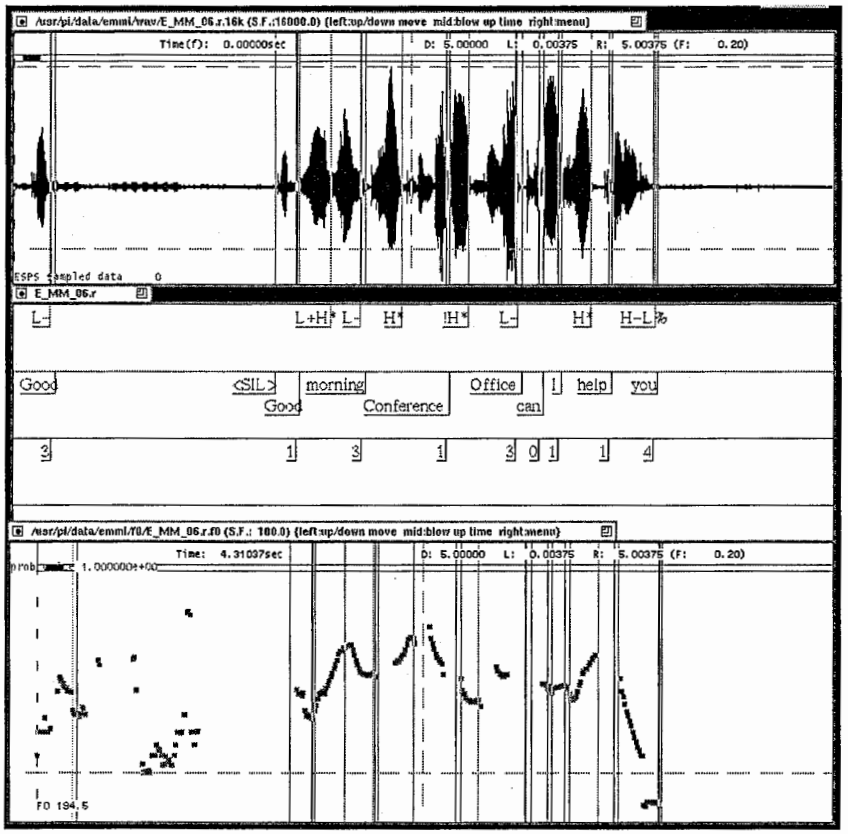
The ten screens in this section are taken from a conversation in the *emmi* data. The data shown here decribes the agents' side of the conversation and the varying labeling types which were assigned to it.

The conversation can be called up using the *xwaves* command with the file name

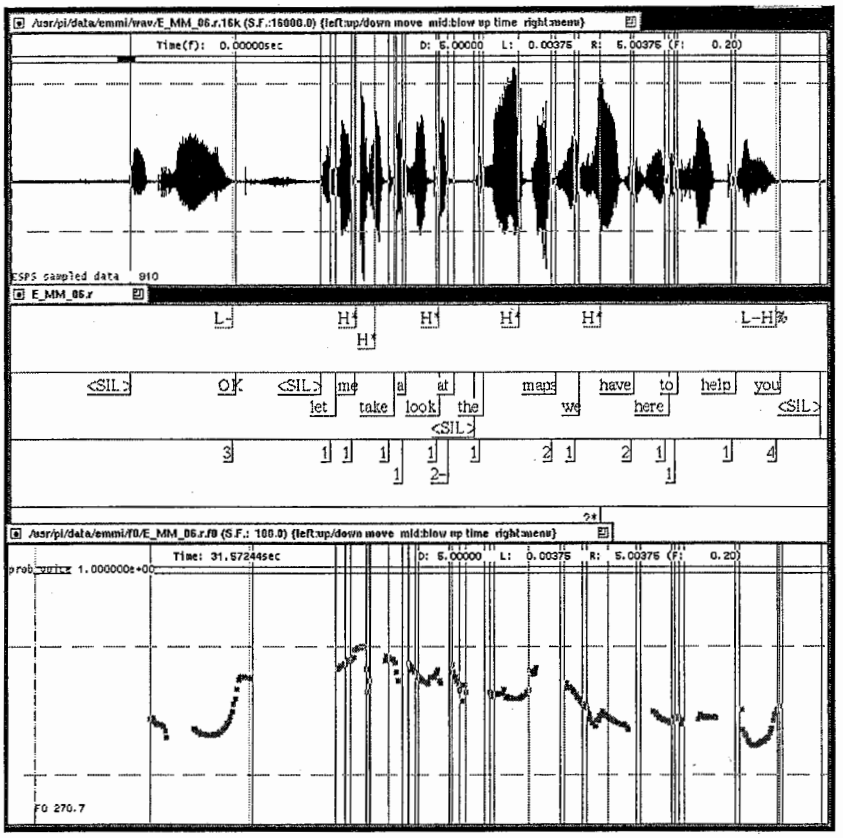
```
E_MM_06.r
```

Directory : /usr/pi/data/emmi/ToBI-lbl

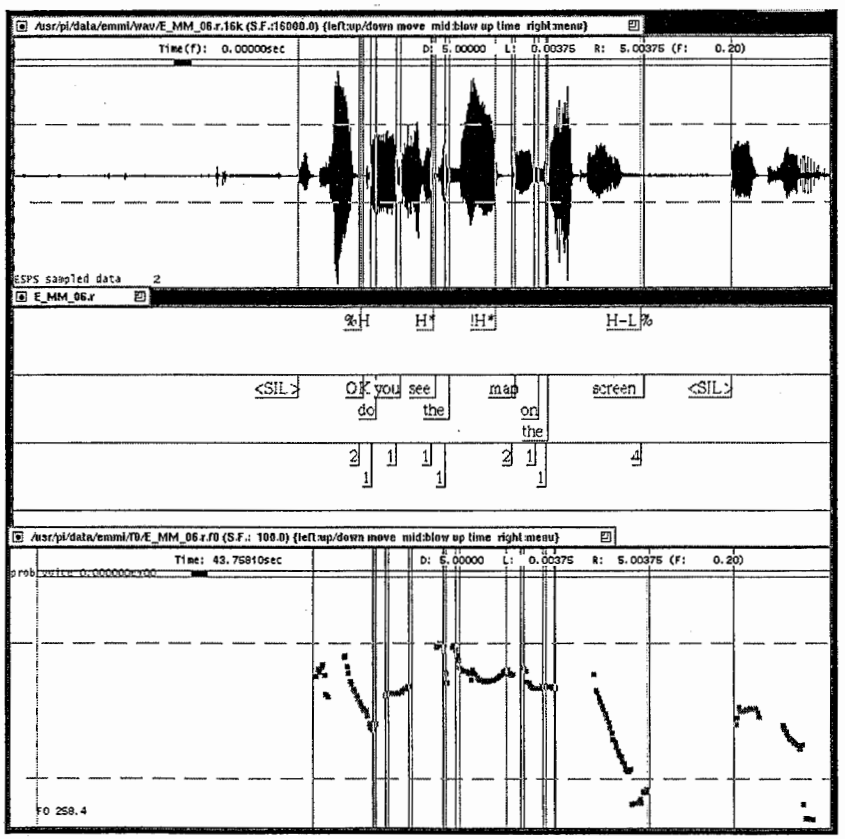
Screen 1



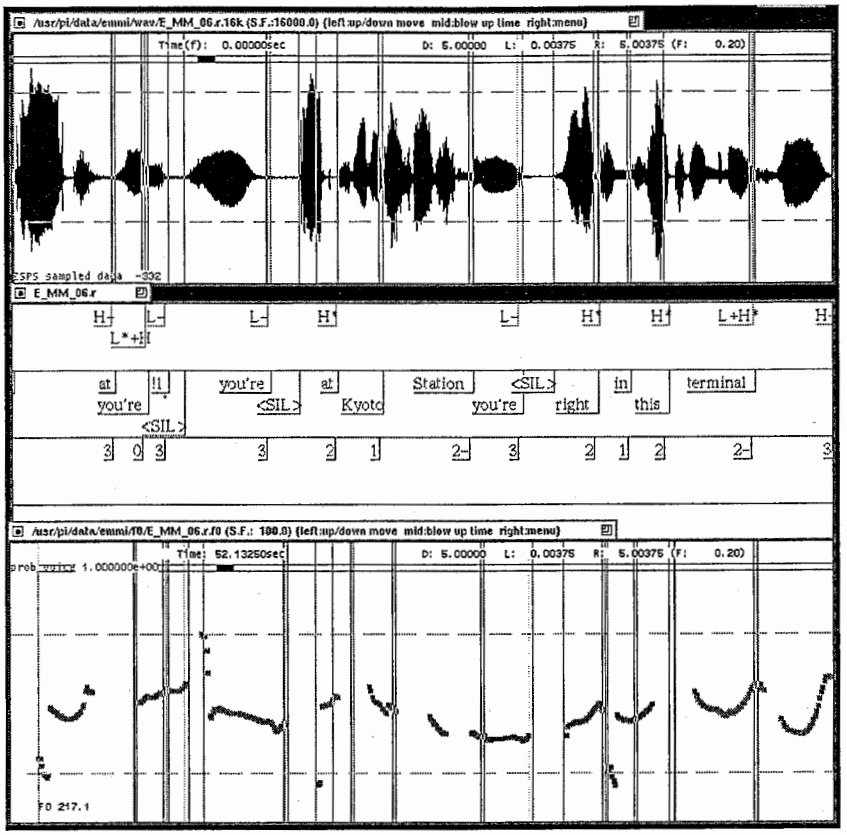
Screen 2



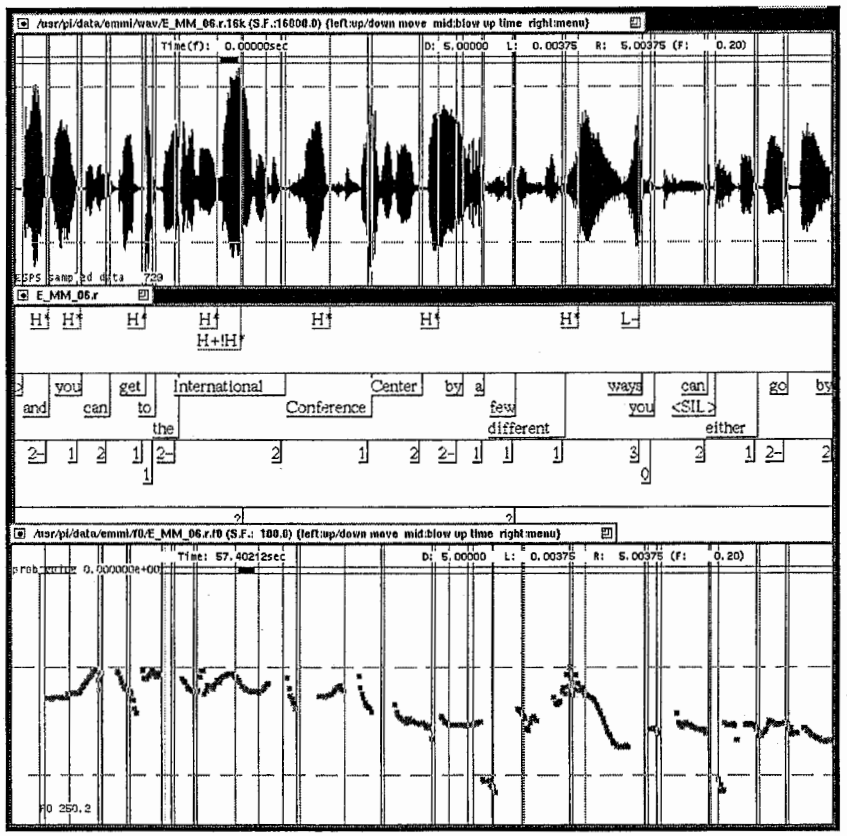
Screen 3



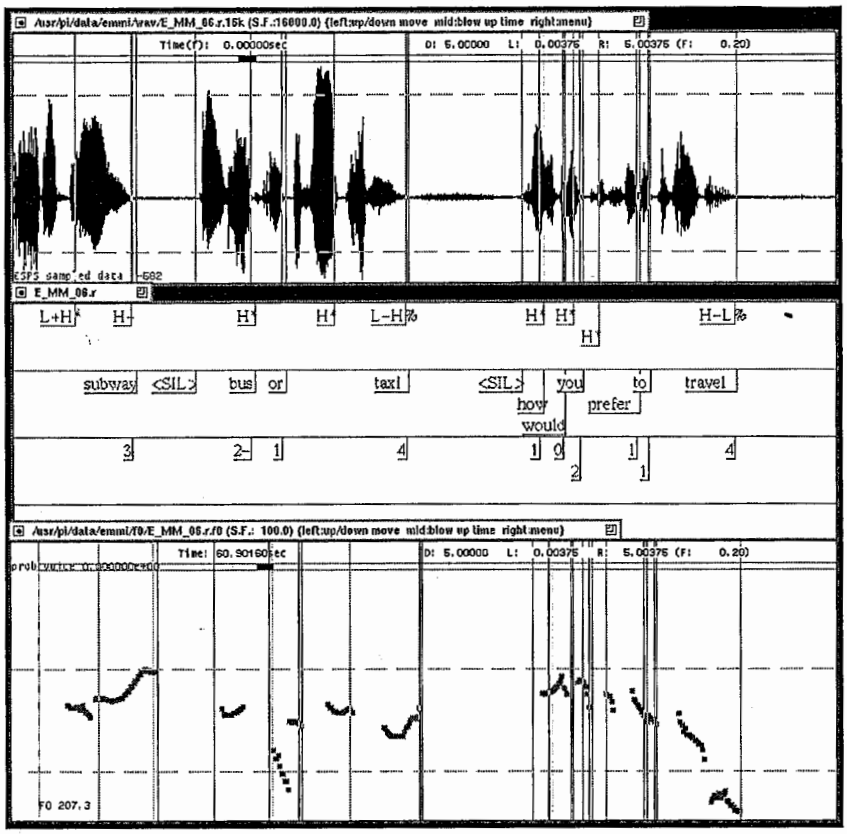
Screen 4



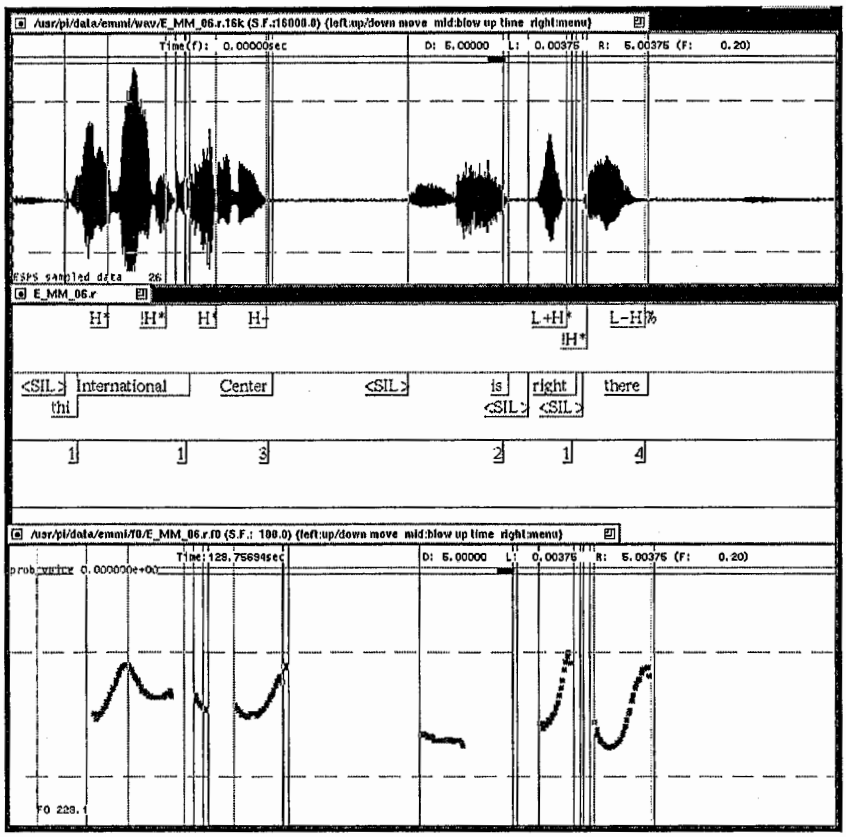
Screen 5



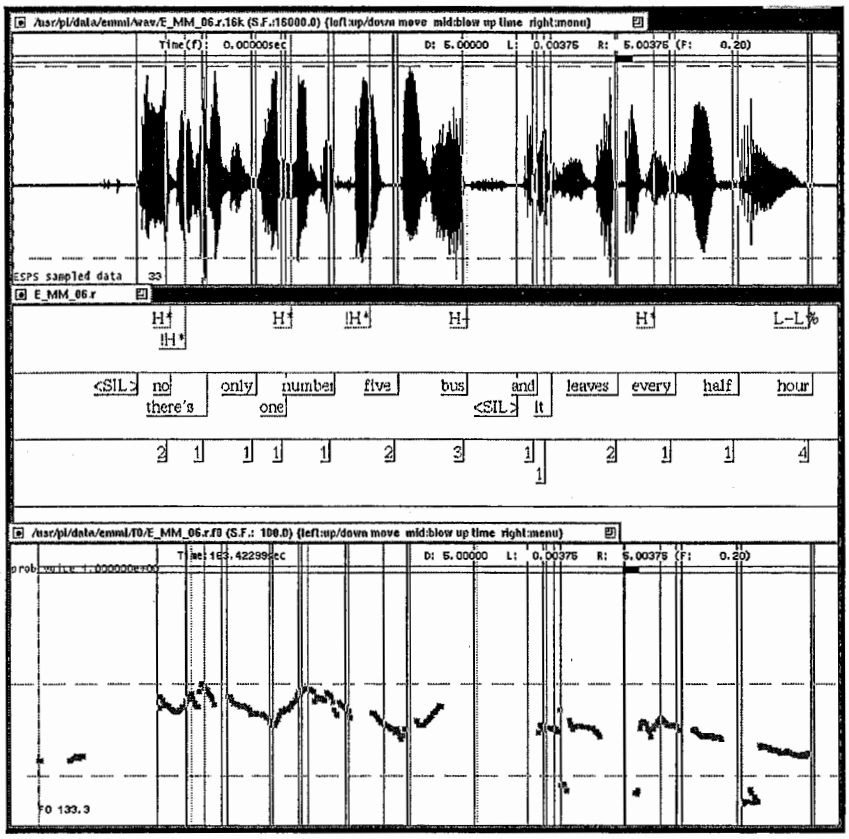
Screen 6



Screen 7



Screen 8



Screen 9