Spontaneous Speech Corpus of Japanese

Kikuo Maekawa*††, Hanae Koiso*, Sadaoki Furui†*, Hitoshi Isahara††*

* The National Language Research Institute
3-9-14 Nishiga'oka, Kita-ku, Tokyo 115-8620 Japan
{kikuo, koiso}@kokken.go.jp
† Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo 152-8552 Japan
furui@furui.cs.titech.ac.jp
†† Communications Research Laboratory
588-2, Iwaoka, Nishi-ku, Kobe 651-2401 Japan
isahara@crl.go.jp

Abstract

Design issues of a spontaneous speech corpus is described. The corpus under compilation will contain 800-1000 hour spontaneously uttered Common Japanese speech and the morphologically annotated transcriptions. Also, segmental and intonation labeling will be provided for a subset of the corpus. The primary application domain of the corpus is speech recognition of spontaneous speech, but we plan to make it useful for natural language processing and phonetic/linguistic studies also.

1. Introduction

It is widely agreed that study of spontaneous speech is a very important but quite difficult research area that should be explored in the near future. Many phoneticians and speech engineers are aware of the fact that 'real' human speech communication is quite different from what they are analyzing in the laboratories, i.e., read-speech.

There are efforts to enlarge the existing knowledge and technologies to the area of spontaneous speech as summarized in Sagisaka, Campbell & Higuchi (1996, See especially the chapter by Beckman). But it seems that all these studies are encountering the same type of obstacle, namely, the quantitative limit of the data.

Study of spontaneous speech requires larger amounts of data than in the study of read speech for at least two reasons.

For one thing, spontaneous speech is inherently more variable than read speech. Different groups of speakers have different manners of speaking depending on social attributes such as age, gender, education, profession, and so on (NLRI, 1953). Moreover, one speaker could speak in quite different speaking styles according to social and personal conditions (Labov, 1972). These inter- and intra-speaker variabilities require analyses based on a large scale database.

For another thing, a speech sample can not be called spontaneous if its linguistic message is completely pre-fixed. Spontaneous speech should be constructed by speakers on site. This means that there is not much room for researchers to make experimental design to reduce the time and cost of data collection.

Accordingly, a large scale corpus of spontaneous speech is desirable as well as necessary for understanding the linguistic nature of spontaneous speech in order to make a breakthrough in the development of speech and natural language processing technologies for application to real human speech.

In this paper, we will describe the basic design and the status quo of a spontaneous speech corpus that we will compile as a main product of a five-year national project conducted jointly by the National Language Research Institute and Communications Research Laboratory. The project, entitled *Spontaneous Speech: Corpus and Processing Technology* is supported by a grant from the Science and Technology Agency and supervised by the third author of this paper.

Among the three goals of this project, namely, (1) compilation of a large scale spontaneous speech corpus, (2) investigation and modeling of spontaneous speech, and (3) investigation of spontaneous speech recognition and summarization technology, the second and the third depend heavily upon the completion of the first.

2. Corpus Design

2.1. Corpus Size

It is very important to have a clear-cut view of the application when we start compiling a corpus. In our project, we will use the corpus mainly for two purposes, 1) Construction of the language model for speech recognition for spontaneous speech, and 2) linguistic-phonetic and/or natural language processing studies of spontaneous speech.

There is, however, a trade-off between the two purposes: the former requires a large amount of data, while the latter puts more emphasis upon the accuracy and quality of annotations rather than the corpus size. Enlargement of the corpus size and refinement of annotation both result in increased cost of compilation. We try to avoid this problem by applying different annotation strategies to the whole corpus and a substantially smaller subset of it.

As shown in table 1, the whole corpus will contain 7,000,000 morphemes. Based on our prior experiences, we regarded this to be the minimum requirement for a language model for a new speech recognition system. Digitized speech, its transcriptions, and morphological annotations of the transcribed speech will be the contents of the whole corpus.

On the other hand, 500,000 morphemes out of the total of 7,000,000 will be in the smaller corpus called the *Core*, to which we will concentrate the cost of annotation. The Core will be provided with segmental and intonation

labeling in addition to the above-mentioned contents. The strategy of segmental and intonation labeling is discussed in section 4.2 below.

	Amount of Data	Contents
Whole corpus	7,000,000 morphemes (800-1000 hours)	Digitized speech Transcription Morphological annotation
Core	500,000 morphemes (55-70 hours)	Segmental labeling Intonation labeling (in addition to the above)

Table 1. Corpus Size and Contents

The core has one more *raison d'être* in our project. Although we will provide morphological annotation, namely word boundary and part of speech tagging, for the whole corpus, it is extremely difficult to annotate the corpus of this size by hand. So, we plan to use the Core as the learning data for the automatic morphological analysis software that we are developing in our project, with which the whole corpus will be annotated. On the other hand, morphological analyses of the Core will be done manually by the lexicographers at the National Language Research Institute (NLRI).

2.2. Language Variety

Selection of speech variety is another vital issue of corpus design. Since it is practically impossible to make a corpus that covers all varieties of a given language, it is desirable as well as necessary to concentrate upon a specific variety.

In our project, we concentrate upon spontaneous monologue rather than dialogue, because, modeling of dialogue speech for speech recognition requires quite different approach than that of monologue, and we think we can not handle both of them satisfactorily in a five-year project.

Also, we concentrate upon a social variety called *Common Japanese* (CJ). In today's Japan, people who are educated at least in high school, speak two varieties: their native dialect and the CJ. The later is a variety used in more or less formal situations like business/academic meetings or public lectures in front of an audience.

The segmental phonology, syntax, and lexicon of CJ spoken by people who are in their fifties or younger are quite similar to those of Tokyo Japanese. Lexical accent, however, differs considerably reflecting the phonology of the speakers' native dialects even among younger speakers.

So, we make it our principle not to pay attention to prosody and concentrate our attention on segmental and syntactic characteristics. A speech is classified as CJ, and then stored in our corpus, if its segmental, syntactic and lexical characteristics approximate those of Tokyo Japanese.

According to our pilot evaluation, out of the total of 289 speeches that we have recorded in the past six months, 278 were evaluated as CJ.

2.3. Sources

Because it is our intention to make the corpus open for researchers, all speech materials in our corpus must be copyright-cleared. We make it a rule to make recordings of those speakers who agreed, with written consent, to provide their speech for our corpus, thereby making it open for academic use.

Currently, we are making two different types of data recording: *academic presentations* (AP) and *simulated public speech* (SPS).

By AP is meant the live recording of researchers' presentations in various academic meetings and we plan to record at least 300 hours of academic presentations.

Currently, our recordings are limited to speech related academic societies like The Acoustical Society of Japan, The Phonetic Society of Japan, and The Society for the Study of Japanese Language, but we are planning to correct this bias by enlarging the number of societies.

Also, it is important to note that the distributions of speakers' age and sex are strongly skewed in AP data. Most speakers are male and in their twenties or thirties.

On the other hand, SPS is short speech (mostly 10 to 15 minutes long) spoken specifically for our corpus by paid non-professional speakers. They are instructed to prepare an outline of their talk instead of the completely pre-fixed text. SPS is recorded mostly in the recording studio of the NLRI in front of a small audience. The topic of the speech can vary from speaker to speaker.

We plan to make at least 450 hours of SPS in which distributions of speakers in terms of their age and sex are maximally balanced.

At this point, readers might want to know if samples from broadcasting sources are involved in the corpus. We had negotiations with some broadcasting companies, but, unfortunately, it turned out that broadcasting companies are generally very unwilling to clear the copyright of their sources . So, most probably, broadcasting sources will not be involved in our corpus.

We could make recordings of about 70 hours of AP and 70 hours of SPS during the past six months, since the beginning of our project.

2.4. Degree of Spontaneity

When we started data collection, we made it a principle to exclude speech that were completely pre-fixed. As we went on with live recordings of AP, however, it turned out that spontaneity of speech could vary even within one presentation.

For example, spontaneity was reduced considerably when a talker was reading a manuscript prepared for the presentation, but even in this type of well-prepared presentation, degree of spontaneity was not always low throughout the presentation. It suddenly increased when the talker made a digression or found mistakes in the manuscript. Judging from this experience, now we do not pay much attention to the preparedness of the presentation.

Consequently, our corpus involves samples that hardly will be classified as spontaneous speech if we interpret the term very rigorously. Rough estimation suggests that about 5 percent of the whole corpus will be of this sort.

We believe, however, inclusion of prepared speech does not deteriorate our corpus. Because, for one thing, even the least spontaneous samples recorded so far are distinctively more spontaneous than the typical read speech, such as professional announcers' news reading, because the prepared speech in our corpus contain many

fillers disfluencies.

For another thing, spontaneity is a matter of more or less, and we need a wide range of samples differing in spontaneity in order to know what really are the phonetic and/or linguistic characteristics of speech spontaneity. To help conduct this sort of study, all speech samples in our corpus will be provided with subjectively evaluated indices of spontaneity, ranging from 0 (=completely prepared-to-be-read) to 5 (= completely spontaneous).

Table 2 is the tentative list of information that will be provided for each sample.

Speaker Information	Age and sex Birth place and past and present living places. Education level Existence of prepared manuscript etc.
Speech and Recording Information	Recording date and place Recording equipment used Degree of spontaneity Evaluation of fluency Evaluation of proficiency Description of voice quality characteristics (if any) Description of noise sources (if any) etc.

Table 2. Information about speakers and speech samples that will be involved in the corpus

2.5. Recording

Speech samples are recorded using unidirectional head-worn condenser microphone and digital tape recorder (DAT) in 48kHz sampling frequency and 16 bit quantization. All samples are downsampled to 16kHz before being stored in the corpus.

Although maximum care is taken to make good recordings, the recording condition varies considerably from one recording to another in live recordings of AP, due mainly to the difference of the room acoustics of conference sites.

The condition of SPS is generally much better than that of AP , because they are recorded mostly in the recording studio of the NLRI.

Video recordings are also made. Videos can be quite helpful in the speech transcription work, because we can get much information about the content of speech by checking the viewgraphs used in presentations.

They are also useful to know the reasons of sudden interruptions (floor littered with viewgraphs or sudden-death of a presentation computer) or the interaction between the speakers and the chairperson who may intervene in the presentation.

Unfortunately, however, recorded videos will not be involved in our corpus, because the usage of videos is restricted to the transcription work in the copyright-clearance contract. People are generally very reluctant to give consent to open usage of their video images.

3. Transcriptions

3.1 Orthographic and Phonetic Transcriptions

Figure 1 shows an example of transcription, whose format is fairly complex. This format was devised to satisfy as much as possible the requirements both from speech recognition and natural language processing studies, which can be incompatible at times.

Our transcription file contains two different kinds of transcriptions in a line separated by an ampersand. In the left-hand column is the transcription that we call *orthographic*, which is written using both Kanji (Chinese

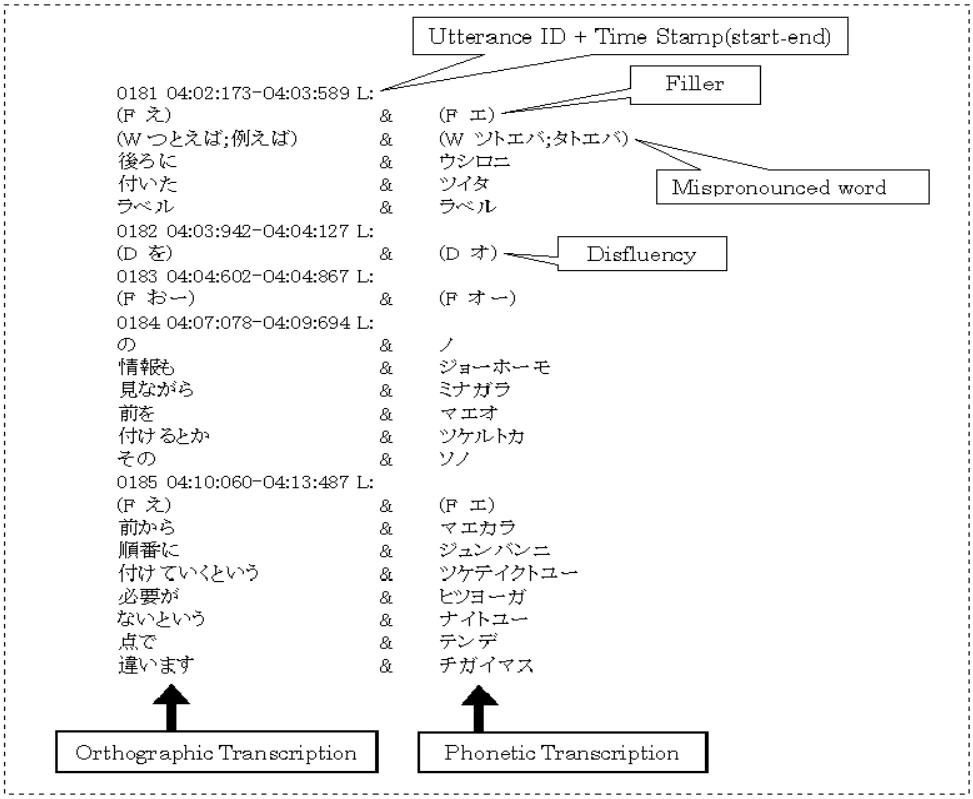

Figure 1: Example of transcription

logographs) and Kana (Japanese syllabary). Transcription in the right-hand column is called *phonetic* and is written in Kana only.

Orthographic transcription will be used in language modeling for speech recognition. It will be used also in automatic morphological analyses; morphological analysis programs for Japanese text require that input text be written in Kanji and Kana because the boundary between Kanji and Kana provides precious information about word boundary (Japanese orthography does not use blank spaces to indicate word boundary).

Though we call it orthographic, our orthographic transcription is different from the standard orthography of Japanese in some respects. Most importantly, our orthographic transcription does not allow any variation of word-to-letter correspondence that characterizes the standard orthography

For example, Japanese word *uchiawase* (meeting) can be written at least in six different ways shown in figure 2.

打ち合わせ　打ち合せ

打合わせ　　打合せ

打ちあわせ　うちあわせ

Figure 2: Example of Orthographic Variation

We try to exclude this sort of free variations in writing as much as possible in order to make data search easy and reliable, thereby maximizing accuracy and reliability of language modeling.

On the other hand, phonetic transcription is needed basically to show the readings of Kanji strings (which can have more than two readings very often). At the same time, this transcription shows, as precisely as possible within the limit of syllable letters, the actual pronunciation as it appears in the recorded speech. There are two different cases where we need phonetic transcription.

In Japanese, phoneme-to-kana correspondence was considerably simplified after the WWII, but there are some residues of historical orthography where letters and real pronunciation diverge. Well-known examples are grammatical particles and long vowels.

Also, phonetic transcriptions reflect those segmental variations that characterize spontaneous speech, of which we quote only two well-known examples.

The first example is the variation of phonologically long vowels. The loanword from English, “computer”, can be pronounced either as [kompju:ta:] or [kompju:ta] in IPA notation, the difference consisting in the reduction of a long vowels in the word final position. Any loanword from English ending with a long vowel can be subject to this variation.

The second example is the deletion of the consonant /w/ in word medial position. When /w/ is preceded and followed by the vowel /a/, the consonant is deleted and the resulting vowel sequence /aa/ is pronounced as a single long vowel. The surname of the first author of the present paper, /maekawa/, for instance, is often pronounced as [maeka:].

3.2. Utterance Boundary

In some lines of figure 1, we see the utterance ID followed by the time stamps showing the beginning and end of each utterance like ‘0181 04:02:173-04:03:589’.

Definition, and recognition thereby, of the utterance unit is certainly one of the most difficult tasks in the compilation of a spoken language corpus.

We make it the fundamental principle to put automatically an utterance boundary at the place where a pause of 200ms or longer emerges in the recording.

We also put utterance boundaries at the places where a short pause (shorter than 200ms but longer than 50ms) follows the typical sentence-ending forms of predicate (verbs, adjectives, and copula). Prosody often plays a crucial role in this task, because the sentence-ending form of a predicate is identical to its adnominal form in present-day Japanese. Final lengthening and/or existence of a boundary tone can be good indicators of an utterance boundary.

SYMBOL	MEANING
(F)	Fillers.
(D)	Disfluency.
(W;)	Mispronounced word. Supposed-to-be correct form is shown after the semicolon.
(?,)	Uncertainty in perception. Multiple candidate words are shown separated by comma if necessary.
(M)	Metalinguistic expression
(Laugh) (Cough) (Yawn)	Non-verbal vocal events that co-occur with speech such as laughter, cough, and yawn.
<Laugh> <Cough> <Breath> <Lip>	Non-verbal vocal events that do not co-occur with speech. <Lip> means lip noise.
<P>	Pause.
<H>	Prolonged word-final vowel that functions as a filler.
<FV>	Uncertainty of phonetic quality of vowels used as filler.

Table 3: Tentative list of tags used in transcription

3.3. Filler, Disfluency and Noise

Fillers, or filled pauses, are among the most eminent indicators of spontaneity of speech. They are marked by a tag (F) in both orthographic and phonetic transcriptions (see figure 1). A typical filler is a prolonged monophthong accompanied with flat pitch, but short conjunctions like /ano/ and /sorede/ are recognized as fillers when they are accompanied, typically, by prolongation of the last vowels and flat pitch

Disfluency, another indicator of spontaneity, is marked using tags (D) or (W). (D) is used typically to mark cases where speakers pronounced a word, or fragment of it, and corrected it later, while (W) marks the cases without correction, *i.e.* cases where speakers are not aware of the mispronunciation. Table 3 shows the tentative list of tags used in our corpus.

4. Annotations

As noted earlier, we have two different schemata of annotation for the whole corpus and the Core. These schemes will be described in the following subsections.

4.1. Morphological Annotation

As noted earlier, all speech samples in our corpus will be analyzed in terms of word boundaries and parts of speech.

A big problem of morphological analysis is that there is no widely-agreed-upon definition of word in Japanese. This is partly because Japanese orthography does not have the custom of showing word boundaries by blank spaces, but more fundamentally, this is a reflection of the linguistic characteristics of Japanese morphology which allow quite free word-formation.

For example, the name of the institution to which the first two authors of the current paper belong, *Kokuritsu kokugokenkyusho*, is a long compound noun. But it can be broken down into at least three elements, *kokuritsu* (adjective 'National'), *kokugo* (noun 'National language'), and *kenkyuusho* (noun, 'Laboratories'); all three elements having the full status of word. Furthermore, the last word can be broken down into *kenkyuu* (noun, 'research') and *sho* ('institution'), the last one being an element smaller than a full word, i.e., morpheme. This example shows that, in Japanese, "word" is a theory-dependent entity.

Moreover, criteria of word recognition can be different depending on the purposes of morphological analyses. This makes our analyses more complicated.

As a general tendency, speech recognition prefers to recognize longer units as word while natural language processing and linguistics prefer shorter units.

In our morphological analysis, we plan to reconcile this problem by providing two different analyses based on two different working definitions of word, namely longer and shorter words.

In the longer word analysis, the example cited above, *Kokuritsukokugokenkyuusho*, is analyzed as a single word unit, while in shorter word analysis, the same string will be broken into four word units as *kokuritsu* + *kokugo* + *kenkyuu* + *sho*. Roughly speaking, these two units correspond to the units that have been utilized in the lexicostatistical surveys conducted by the NLRI (NLRI, 1983, among many others).

4.2. Annotation of the Core

As mentioned earlier, annotation of the Core will involve segmental and intonation labeling in addition to the morphological annotation.

At the present, we have not started the labeling of the Core, but some pilot labeling experiments are on the way. In the rest of this section, we will give a brief overview of the problems that we encountered during the course of the pilot labeling.

4.2.1. Segmental labeling

Segmental labeling of spontaneous speech is known to be a difficult task, because cues of segment boundaries reported in acoustic phonetics literature, that are mostly based upon the analyses of read speech, often change their shapes or even disappear in spontaneous speech. Our very preliminary labeling experiment revealed the following tendencies.

1) Lenition of stops: voiced stop consonants are realized often as a sort of [+continuant] consonant in non word-initial positions. This is true with voiceless stops but to a lesser extent.

2) Devoicing of non-close vowels: close vowels, /i/ and /u/, are devoiced almost regularly when preceded and followed by voiceless consonants. In spontaneous speech, non-close vowels are often devoiced in the same phonological contexts.

3) Drop of close vowels: close vowels can disappear in non-devoicing environment, i.e., when they are preceded by a voiceless consonant but followed by a voiced homo-organic consonant: for example, *kokugo* (national language) and *kokokara* (from here) can be pronounced as [kokgo] and [kokkara] respectively.

4) Reduction of /r/: the /r/ consonant is reduced to a considerable extent when it is preceded and followed by same vowels, e.g. /terebi/ (television) is realized as something like [te:bi], the first long vowel being a sort of retroflex vowel.

These should be a small subset of the phonetic and/or morphological variations that characterize spontaneous speech. We plan to show these variations by using a sub-phonemic labeling system. But the extent to which we can pursue the accuracy of labeling is limited by the cost of annotation. Probably it is better to concentrate on those specific variations that are the most interesting.

4.2.2. Intonation labeling

We conducted a pilot intonation labeling using J_ToBI labeling scheme (Venditti,1995; Campbell,1997), which is based upon the theory of Japanese intonation proposed in Pierrehumbert and Beckman (1988).

Although this experiment revealed several problems that suggested the need to revise and/or extend the current scheme, we will discuss only one of them, namely the treatment of boundary tones. See Maekawa and Koiso (2000) for other problems.

In the theory proposed by Pierrehumbert and Beckman, boundary tones are regarded to be the properties of utterance, *i.e.* the highest node in the prosodic tree. This means that boundary pitch movements like question rise can occur only at the end of an utterances. The J_ToBI scheme inherits this hypothesis in that it allows boundary tones only at the boundaries whose strength is marked by a Break Index 3 that covers both the intermediate phrase and utterance boundary in the underlying theory.

In some speakers' spontaneous speech, however, nearly all short syntactic phrases are marked by local pitch movements that look like boundary tones. Figure 3 shows an example of such an utterance, in which a young male graduate student is talking about speech recognition at an Acoustical Society of Japan meeting.

This utterance is consisted of five syntactic phrases, or *bunsetsu*, which are usually realized as accentual phrases in prosody. However, the ends of the second, third, and fourth phrases are marked by locally rising pitch movements (as indicated by circles in the figure) which are quite similar to the authentic boundary tones that occur at the end of an utterance.

It is also important to note that downstep, *i.e.* narrowing of pitch range triggered by the presence of a lexical pitch accent, seems to continue across the seemingly boundary tones.

The problem is that, in the current scheme, presence of a boundary tone requires a Break Index 3 on the one hand, and the presence of BI3 implies the resetting of the effect of downstep. So, if we attach greater importance to the

continuation of downstep, the Break Indices should be 2, that means the end of an accentual phrase.

To sum up, we are encountering a mismatch between the break index indicated by a local pitch movement and that by a global trend in pitch.

The mismatch can be even more complicated when we see the cases where the seemingly boundary tones were followed by short pauses, because, usually, pauses are interpreted to be an indicator of BI 3.

How should we handle these mismatches? Some possible solutions were suggested in Maekawa and Koiso (2000), but we would like to make them an open question for the coming years of our project. Currently, we are planning to start a Japan-U.S. collaboration to extend the J_ToBI scheme for spontaneous speech (See Beckman and Venditti, 2000). The mismatch problems will be one of the central issues of this collaboration work.

5. Prospects

What follow are the temporal landmarks in the schedule of corpus compilation:

- Finish data collection by the end of March 2002.
- Finish morphological annotation of the Core by the end of March 2002.
- Finish transcription by the end of March 2003. And,
- Finish everything by the end of March 2004, which is the end of our project

After finishing the compilation, we plan to make the corpus available free-of-charge for academic purposes. We also plan to make part of the corpus open for monitoring even before the end of compilation.

The corpus, which we tentatively call the *Corpus of Spontaneous Japanese*, will be the world's first of this type, and it is our wish that it is used by many people, both in scientific and technological fields, both at home and abroad. We welcome comments on our project and would like to exchange the ideas and experiences of corpus compilation with colleagues who are working on similar projects.

References

Beckman, M. (1996). A typology of spontaneous speech. In Sagisaka. Y. *et al*. Eds. (1996), 7-26.

Beckman, M. and J. Venditti (2000). Tagging prosody and discourse structure in elicited spontaneous speech. In *Proceedings of the International Symposium: Toward the Realization of Spontaneous Speech Engineering* (pp. 87-98), Tokyo: NLRI and CRL.

Campbell, N. (1997). The ToBI system and its application to Japanese. *Journal of the Acoustical Society of Japan*, 53-3, pp. 223-229.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Maekawa, K. and H. Koiso (2000). Design of a Spontaneous Speech Corpus for Japanese. In *Proceedings of the International Symposium: Toward the Realization of Spontaneous Speech Engineering* (pp. 70-77), Tokyo: NLRI and CRL.

NLRI (1953). *Language Survey in Turuoka City*. Report No.5 of the NLRI. Tokyo: Shuei Shuppan.

NLRI (1983). *Studies on the vocabulary of senior high school textbooks*. Report No 76 of the NLRI. Tokyo: Shuei Shuppan.

Pierrehumbert, J. and M. Beckman (1988). *Japanese Tone Structure*. The MIT Press.

Sagisaka, Y., N. Campbell & N. Higuchi, Eds. (1996). *Computing Prosody: Computational models for processing spontaneous speech*. NY: Springer.

Venditti, J. (1995). *Japanese ToBI labeling guidelines*. (http://www.ling.ohio-state.edu/phonetics/J_ToBI/).

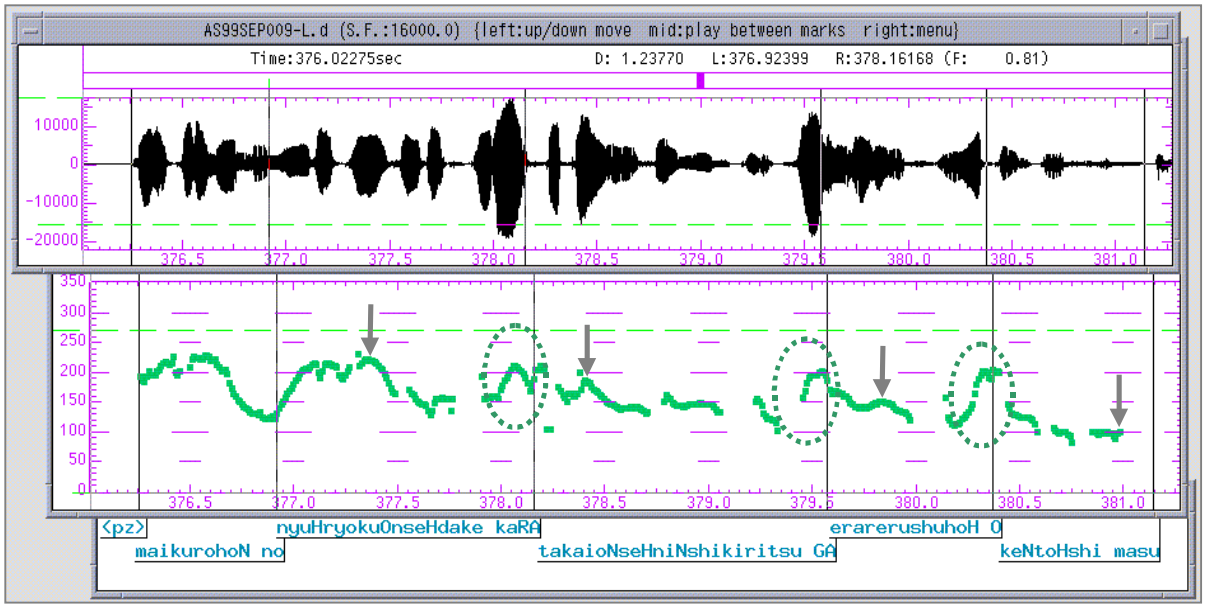

Figure 3: Example of seemingly boundary tones within an utterance.

Waveform (top) and extracted pitch contour (bottom). Vertical lines denote seemingly accentual phrase boundaries. Ellipses at the end of the second, third, and fourth phrases denote the locations of seemingly boundary (rising) tones. Arrows show the locations of pitch peaks that correspond to lexical pitch accents. Note the continuous lowering of accent peaks.