TR-I-0003
TR-A-0004

Japanese Speech Database with Fine Acoustic-Phonetic Transcriptions

多層音韻ラベルをもつ日本語音声データベース

Yoshinori Sagisaka, Kazuya Takeda, Shigeru Katagiri and Hisao Kuwabara

匂坂芳典, 武田一哉, 片桐滋, 桑原尚夫

1987.5

Abstract

A large sized Japanese speech database at *ATR(JSDB-ATR)* is introduced. These speech data are transcribed in multiple ways using acoustic-phonetic symbols for various data access requests and for the convenience of fine acoustic-phonetic analysis. For multiple transcription, three types of categories are considered: linguistic and phonemic categories, acoustic event categories and some alophonic variation categories. These transcriptions were carried out manually by trained labelers using digital sound spectrograms and several acoustic parameters that reflect speech characteristics. To date, about 8500 words respectively uttered by eight professional announcers have been collected with half of them being acoustically-phonetically transcribed.

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

ATR Auditory & Visual Perception Research Laboratories
ATR 視聴覚機構研究所

1. Introduction

Recently,the construction of speech databases has been undertaken in many languages to obtain much needed knowledge for speech recognition, perception and syntheses.[1]-[3] However, there are few Japanese speech database *(JSDB)* large enough for research purposes. Only the *JSDB* developed by *JEIDA* (Japanese Electronics Industry Development Association) is available and it is used to test the performance of speech recognizer. There is no large sized JSDB in which each speech portion is transcribed in terms of acoustic-phonetic symbols.

In this paper, a large sized *JSDB* that is being built at *ATR (JSDB-ATR)* is introduced. In Chapter 2, the contents of speech data to date are shown with their utterance specifications, and recording and digitizing conditions. *JSDB-ATR* speech waveforms are segmented in multiple ways and finely transcribed using acoustic-phonetic symbols which are explained in Chapter 3. In Chapter 4, hand labeled speech data using these symbols are also shown with data on segmentation accuracy.

These multiple transcriptions should prove useful for quick and intelligent access to desired speech portions in speech research.

2. ATR Japanese Speech Database (JSDB-ATR)

2.1 Speech Utterances

Table 1 shows the database to date. It consists of 5229 important words extracted from a Japanese word dictionary, 115 sentences uttered in three different breath groupings and 386 supplementary words. The 115 sentences were extracted from experimental conversations at *ATR* with slight modifications. All of them were uttered by eight (four male and four female) professional announcers and a half of them have been transcribed using acoustic-phonetic symbols.

2.2 Pronouncing Specifications

For the systematic construction of speech database, the first data collection is restricted to standard utterances. These homogenous speech data are intended to be the standard references for the forthcoming ordinary speech data that will have large intra-speaker variance due to amateur and dialectal pronunciation.

To meet these requirements, only professional announcers using the following pronouncing rules were employed. Most of these pronouncing specifications are the same as those used in *NHK* broadcasts (*Japan Broadcasting Corporation*), and are familiar to professional announcers.

(1) Accentuation

Because many Japanese words can be pronounced with the accent in different locations, one accentuation is selected for each word according to *NHK* announcers criteria.

(2) Adoption of long vowel pronunciation

One of the Japanese writing systems (Kana) corresponds closely to the phonemic symbols. Because of this, word pronunciation can be affected by its spelling in some cases, especially in slow-rate speaking or in mora-chopped speaking. For example the word " とういつ(unification) " can be pronounced in

Table 1 Contents of speech data

Item	numbers
Important word	5229
Phonetically balanced word	216
Alphabet	35
Numeric	25
Syllable (native)	101
syllable (non-native)	9
Conversational sentence	115
Total	about 8500 words

two ways : [to:itsu] , [touitsu] . The latter has an exact one-to-one correspondence to its spelling: と /to/, う /u/, い /i/ and つ /tsu/. To prevent useless confusion on acoustic-phonetic transcriptions, the pronunciation is fixed to [o:] (or [e:]) for the words that can be pronounced in two ways: [o:] for [ou] and [e:] for [ei].

(3) Devocalization

In Japanese, the vowels /i/ and /u/ are often devocalized if they are surrounded by voiceless consonants; other vowels are also devocalized in some cases. Devocalization is specified for each word according to the criteria of *NHK* announcers.

(4) Nasalisation

The voiced velar consonant [g] is often nasalized [ng] in Japanese except at the initial position. Almost all internal [g] are pronounced [ng].

(5) Additional phonemic change

Some words which contain difficult phoneme pronunciation strings are converted to conventional pronouncing forms used in *NHK* broadcasting, (e.g. 支出 [shishutsu] (expense) is converted to [shishitsu]).

2.3 Recording and Digitizing Conditions

Figure 1 shows the data collection system block diagram. All speech data are initially recorded onto a PCM encorder-recorder,then anti-aliase filtered (8kHz low pass filter with attenuation of -100dB/oct.), and digitized in 16bit by 20kHz sampling. All these data acquisition and calculation are done on the workstation. For manual labeling, digital sound spectrograms with some acoustic parameters are produced from these data.

3 Acoustic-Phonetic Transcriptions

3.1 Multiple transcriptions of speech data

For the effective use of speech data, some kind of phonemic transcriptions are needed. However, it often becomes a problem where to decide a phonemic boundary in a speech waveform. The difficulties of phonemic boundary decision can be classified into two groups.

One group consists of speech data that have multiple choices of a phonemic boundary. For example, there are two candidates for the boundary between the vowel /o/ and the fricative consonant /sh/ in the utterance /atoshimatsu/ (settlement) denoted ① in *Figure 2 (a)*. The beginning point of the frication /sh/ does not coincide with the end point of the voiced portion of /o/. There are also two

candidates for the boundary between the vowel /a/ and the following silent portion denoted ②,③ in *Figure 2 (a)*. The inconsistent choice of these boundaries can cause many problems in speech measurement and succeeding analysis of speech characteristics.

The other group consists of speech data in which phonemic boundaries are difficult to find. The same utterance /atoshimatsu/ shown in Figure 2 *(a)* contains a devocalized vowel /u/ at the end of the word(④). This portion has acoustically unique characteristics (i.e. there is no spectral change), and it is impossible to separate it into the two phonemic segments /ts/ and /u/ .

These difficulties show that phonemic categories are not sufficient for speech transcription and that finer acoustic event categories need to be introduced. From this point of view, fine multiple transcriptions of speech is

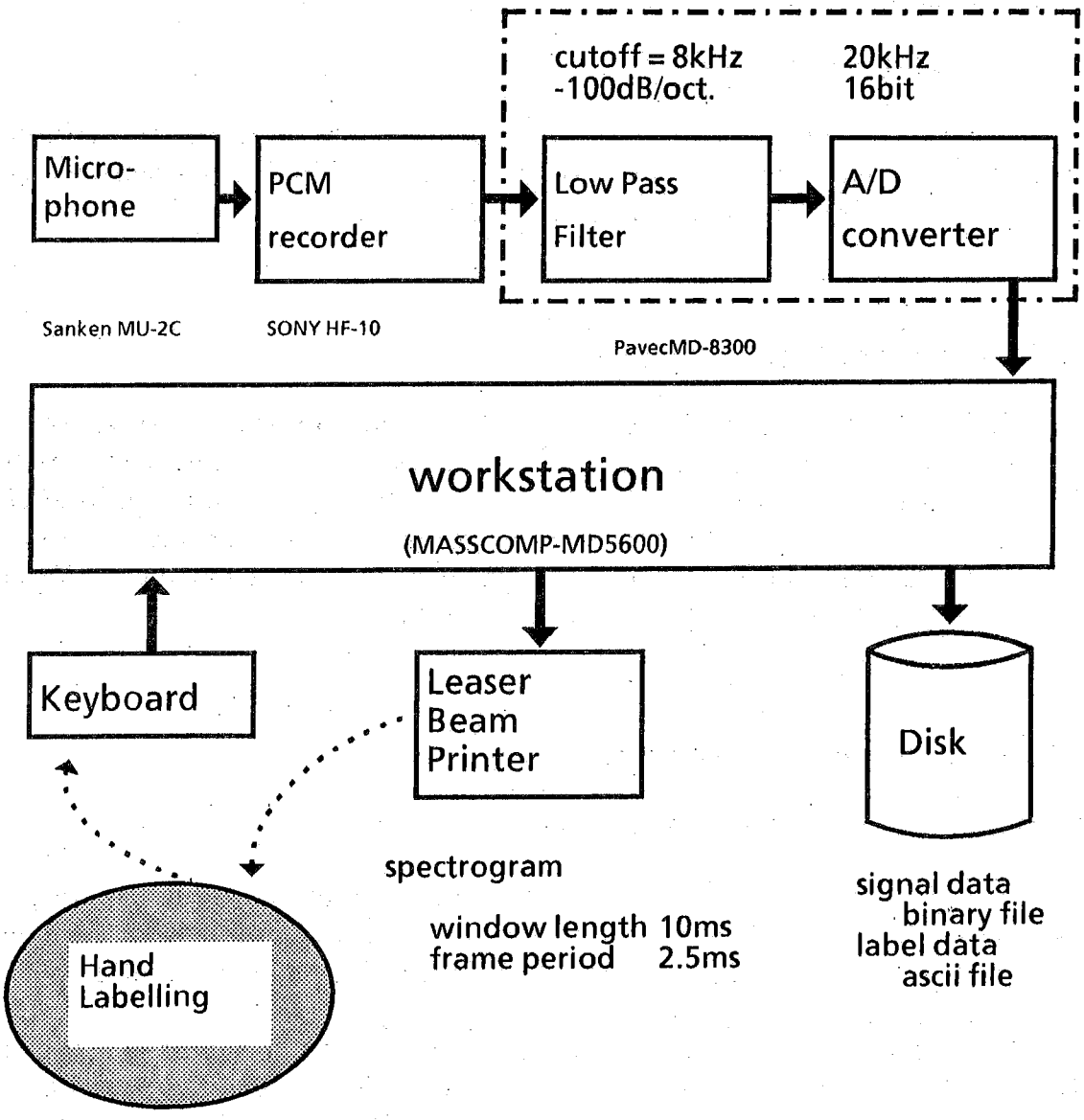

Figure 1 Data acquisition system block diagram

proposed for our *JSDB-ATR*.

Figure 2 (b) shows an example of multiple acoustic-phonetic transcriptions of the utterance /atoshimatsu/ in *JSDB-ATR*. As can be seen in this example, almost all acoustically different portions that can be separated are segmented and transcribed with some symbols. The aim of our transcription is not the precise definition of phoneme boundaries but the fine categorical description of continuous speech. In the following sections, these transcriptions, especially the transcription of acoustic events, are precisely explained.

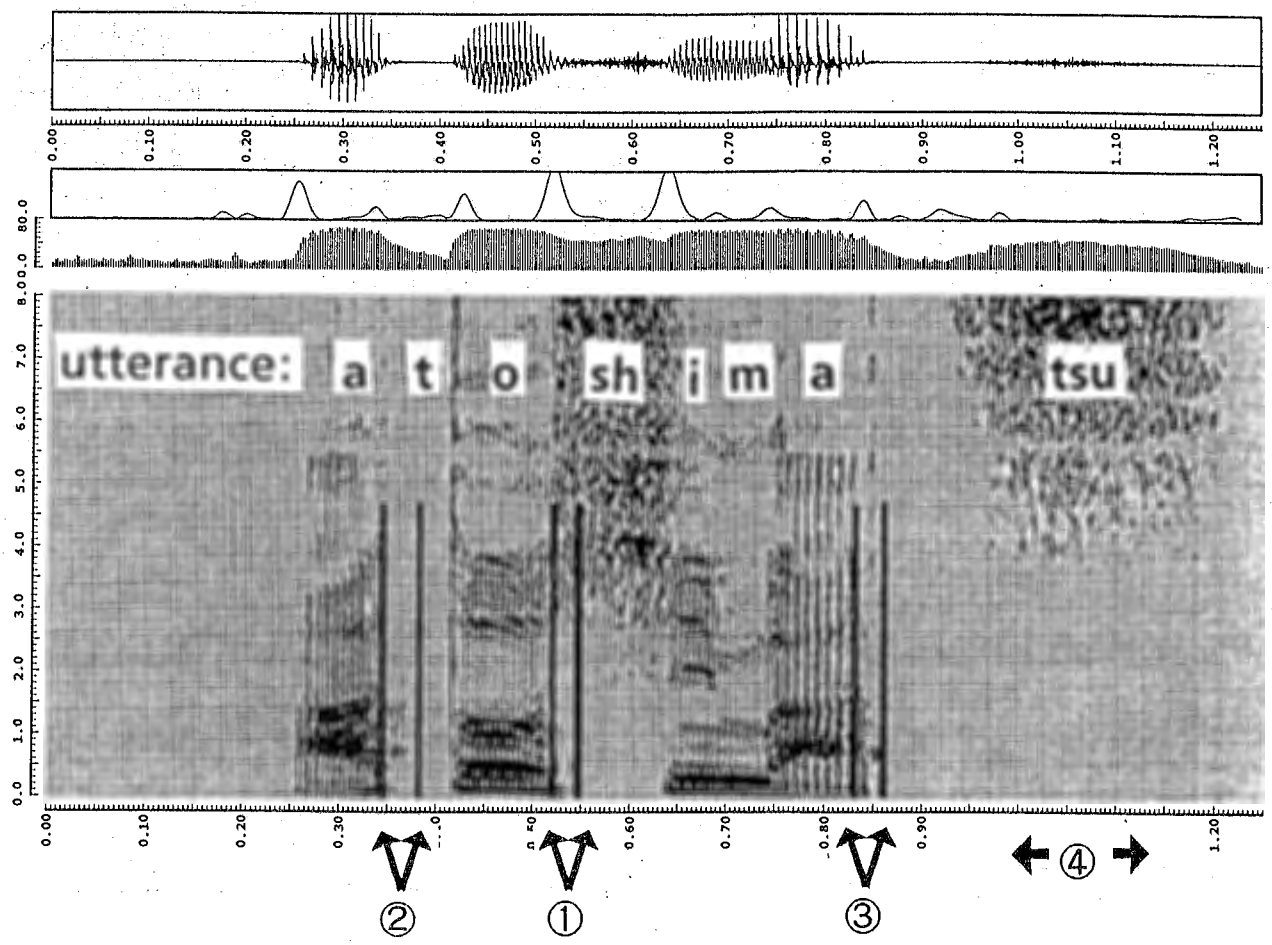

Figure 2 (a) *An example of candidates for the phoneme boundaries (①,②,③) and two inseparable phonemes (④).*

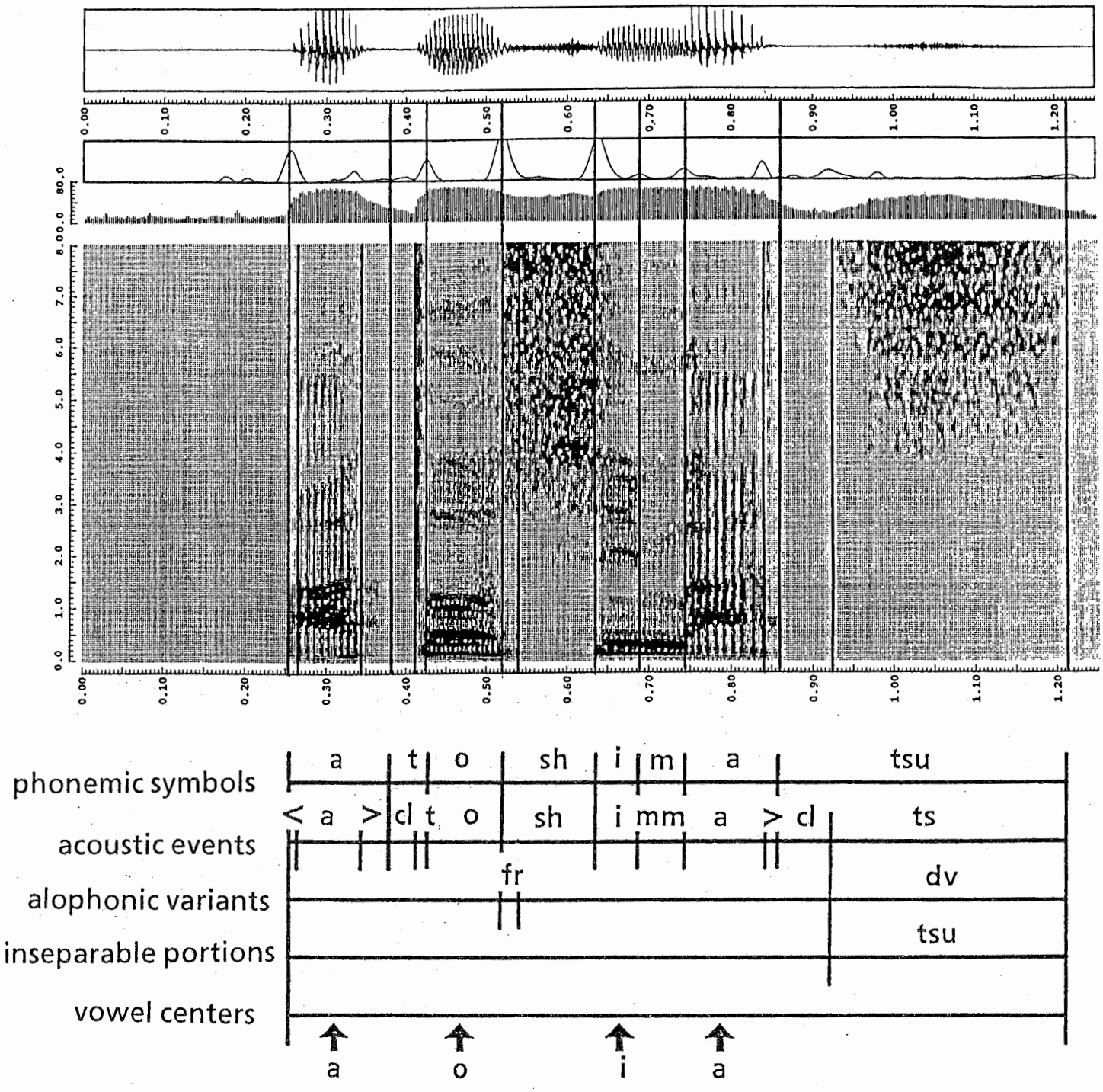

Figure 2 (b) An example of multiple acoustic-phonetic transcriptions of the utterance /atoshimatsu/.

3.2 Linguistic information and phonemic symbols

Linguistic information of the utterance, such as syntactic structure and constituent word attributes are indispensable for same kind of speech analysis. In *JSDB-ATR*, linguistic information of the utterance is to be linked with speech data using an utterance information file. The contents and structure of this utterance information file are now under consideration and not only linguistic information but additional information (i.e. speaker's attributes) will be included. For the present, the word attributes for 5229 word data: multiple word spellings or constituent accents, can be gotten from a Japanese word dictionary.

For phonemic transcription, the Hepburn system of romaji and supplementary expressions for non-native sounds are employed (see Appendix). If these expressions are automatically applied to Japanese word spellings, phonemic symbols are not always the same as phonetic symbols for acoustic events of the corresponding part. *Figure 3* shows the incoincidence of these two symbols. These expressions are adopted for the easy transformation between two different utterances of the same word.

3.3 Symbols for acoustic events

All symbols used for the description of acoustic-phonetic phenomena are summarized in *Table 2*. Though most of these symbols are the same as those used in phonemic transcription, their corresponding portions are slightly different in many cases. These symbols and the criteria for their boundary location have been fixed through many small changes to decrease the deviation of their boundary locations and to hasten labeling speed without losing fineness and consistency of description.

(1) a,i,u,e,o — These symbols stand for the acoustically stable portions of Japanese cardinal vowels. They form the body of corresponding phonemic portions and are transcribed by the same symbols.

(2) <,>,*>,tr — As shown at the vowel locations in *Figure 2 (b)*. and *Figure 3*, the acoustically transcribed portions differ from phonemic ones at the beginning and (or) at the end where the vowel formant structure is not clear. "<" (">") is employed for the vowel beginning (end) portion at the initial (final) word position or at the transition from (into) a voiceless consonant. "*>" is used for vowel end followed by a voiced consonant as shown in *Figure 4*. "tr" is prepared for inexplicable portions in the speech and it is rarely used in normally uttered samples.

The acoustic event transcription for vowels is

schematically shown using these symbols in *Figure 5 (a)*.

(3) p,t,k,b,d,g

In plosive consonants, burst, frication and aspiration portions are merged into one of these symbols. As a matter of course, the separation of these three portions was proposed, but selected portions in Japanese are too short to cut and too weak to distinguish, especially in the voiced consonants (b,d,g). There are several examples containing plosive consonants in which any burst, frication, or aspiration cannot be found, as

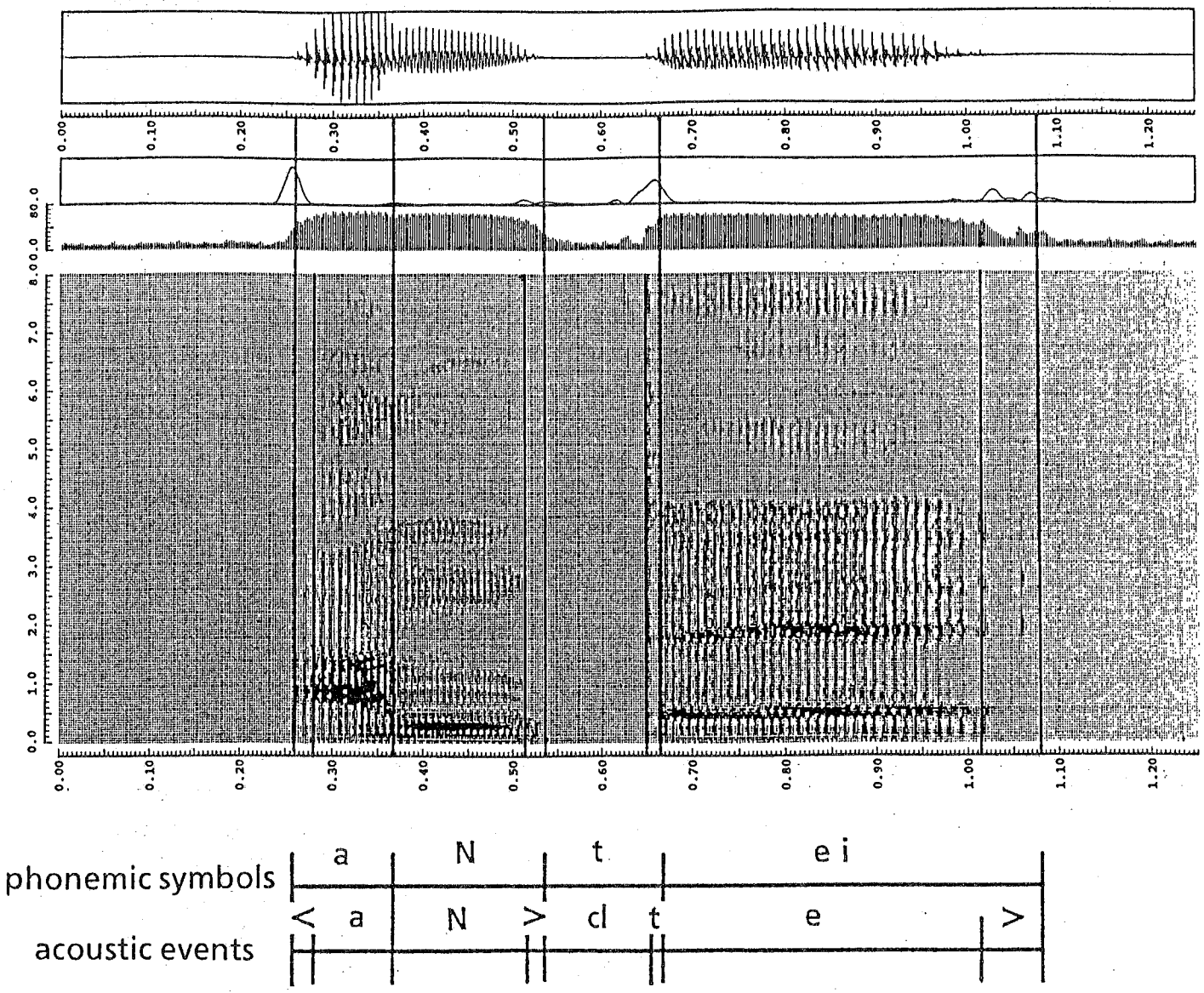

Figure 3 *An example of the incoincidence of phonemic symbols "ei" and the acoustic event symbol "e" in the utterance /aNtei/ (stable). (N: syllabic nasal)*

shown by the schematic pattern for the plosive /g/ in *Figure 5 (b)*.

(4) cl,*cl,mm — For closure portions of voiceless consonants (i.e.silent), voiced consonants (i.e. buzz) and nasals (i.e. murmur), these symbols are used respectively. The symbol "cl" is also used for other silent portions (not for pause intervals) that can be seen in some syllable boundaries or in front of voiced plosives of some speakers.

(5) ts,ch,s,h,sh,z,dj,f — Frication parts are transcribed using these symbols; "ts" and "ch" are for affricates and the others for

Table 2 Symbols for acoustic events

<table>
<tr><th>symbols</th><th colspan="3">acoustic events</th></tr>
<tr><td>a,i,u,e,o</td><td colspan="3">vowel steady portion</td></tr>
<tr><td><</td><td rowspan="3">vowel portion</td><td>preceded by</td><td rowspan="2">voiceless consonant</td></tr>
<tr><td>></td><td rowspan="2">followed by</td></tr>
<tr><td>*></td><td>voiced consonant</td></tr>
<tr><td>tr</td><td colspan="3">phonetically inexplicable portion</td></tr>
<tr><td>p,t,k,b,d,g</td><td colspan="3">burst,frication and aspiration for plosives</td></tr>
<tr><td>cl</td><td rowspan="3">closure for</td><td colspan="2">voiceless consonants (silent)</td></tr>
<tr><td>*cl</td><td colspan="2">voiced consonants (buzz)</td></tr>
<tr><td>mm</td><td colspan="2">nasal (murmur)</td></tr>
<tr><td>ts,ch</td><td rowspan="2">frication portion for</td><td colspan="2">affricates</td></tr>
<tr><td>s,h,sh,z,dj,f</td><td colspan="2">fricatives</td></tr>
<tr><td>r</td><td>liquid</td><td rowspan="2">(coincide with phonemic segments)</td></tr>
<tr><td>w,y</td><td>semi-vowel</td></tr>
<tr><td>N</td><td colspan="3">syllabic nasal</td></tr>
<tr><td>j</td><td colspan="3">palatalized vowel like portion</td></tr>
<tr><td>pau</td><td colspan="3">pause interval</td></tr>
</table>

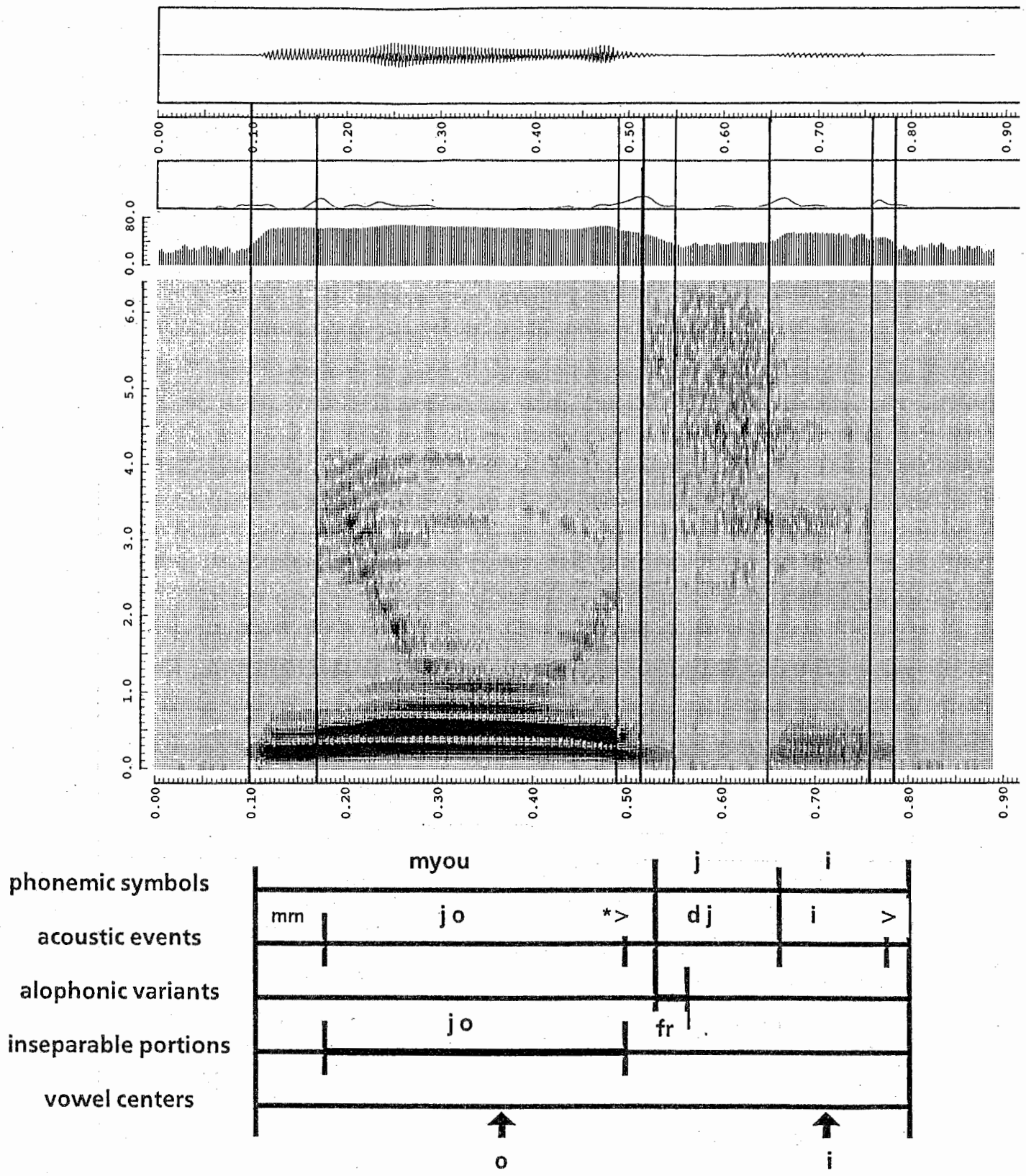

Figure 4 An example of multiple acoustic-phonetic transcriptions of the utterance /myouji/(patronymic)

fricatives. The symbol is aligned at the point where the fricative noise spectrum in the high frequency domain starts as shown in "sh" and "ts" of *Figure 2 (b)* and in "dj" of *Figure 4*.

(6) r,w,y

Liquid "r" and the semi-vowels "w", "y" coincide with the phonemic segments.

(7) N,j,pau

"N" and "j" can be considered as vowel-like portions. As the syllabic nasal /N/ must be preceded by a vowel, phonemic /N/ segments can only contain "N","<" and "*<" as shown in *Figure 3*.

The symbol "j" denotes the /i/, vowel-like portion, that can be found at the beginning part of vowels preceded by palatalized consonants as shown in the /myo/-part of *Figure 4*. Although "j"-portions vowel-like characteristics, they are contained in a palatalized consonant only for the consistency of Japanese consonant notations. Thus acoustic event portions for palatalized consonants can be schematized as (c) in *Figure 5*.

A pause interval is transcribed by "pau".

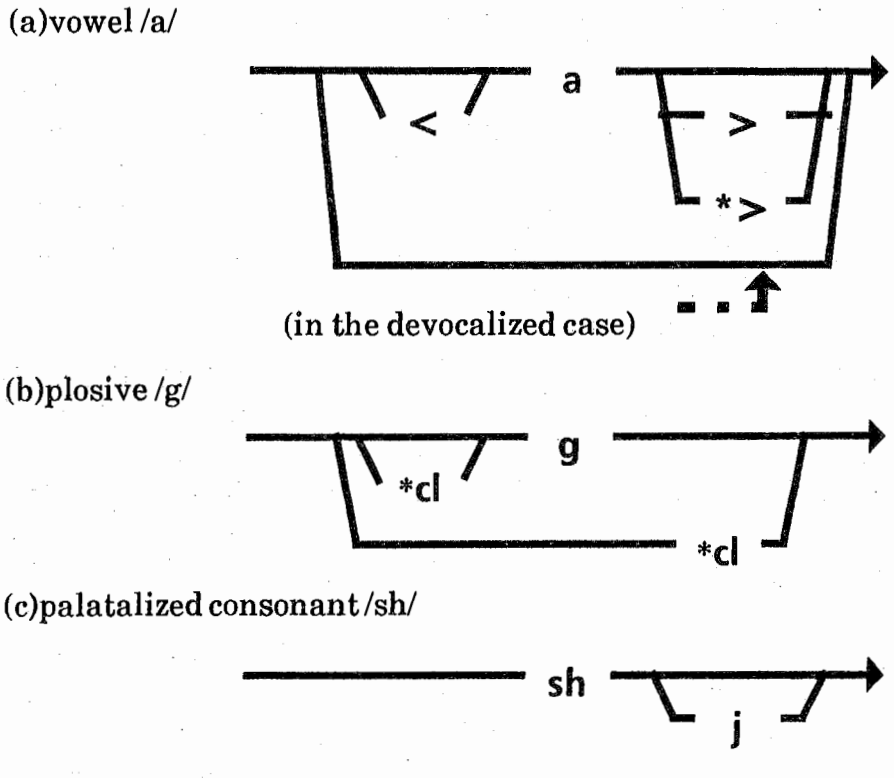

Figure 5 Schematic patterns for acoustic event transcriptions

3.4 Description of alophonic variants

In general, acoustically different portions are differently transcribed using acoustic event symbols. However, there are some acoustic phenomena which usually but not always coincide with one acoustic event. At present, two of these phenomena are transcribed as alophonic variants.

One is a devocalization phenomenon denoted by "dv" and the other is a vowel portion affected by the following fricative consonant and is denoted by "fr". These two phenomena can be seen in *Figure 2 (b)*. Although nasalization was also proposed as an alophonic variants candidate, it was eliminated because the appropriate criteria could not be found for precise positioning.

3.5 Inseparable portions and vowel centers

Throughout the transcriptions of *JSDB-ATR*, segmentation of each transcription into constituents was not impelled where manifest acoustic change could not be found. For this reason, inseparable portions are denoted by multiple symbols. To prevent useless confusion resulting from this notation and to easily access inseparable portions, these portions are specially marked "tsu" as shown in *Figure 2 (b)*.

In addition, the vowel center position is pointed for each vowel as shown *Figure 2 (b)*, *3* and *4*.

4 Hand labeling for acoustic-phonetic transcriptions

Acoustic-phonetic transcriptions in *JSDB-ATR* are done manually by trained labelers. At the present technology level, it is difficult to automatically transcribe speech data using both global and fine analysis. Moreover, at times the automatically chosen boundaries greatly depend on the acoustic parameters employed in the programs and do not coincide with the proper boundaries. For these reasons, hand labeling was chosen as the first method with semi-automatic labeling taken in the second step.[4]

In hand labeling, the biggest problem seems to be labeling segmentation criteria inconsistency. These criteria can fluctuated within one labeler and greatly differ between labelers without manifest cues for segmentation. To decrease the segmentation fluctuation, the criteria have been throughly discussed among specialists and labelers and slightly modified in some cases. As a result, some segmentation criteria were explicitly written in the form of a labeling manual for ambiguous samples.

According to a statistic analysis of the hand labeling accuracy of our labelers, more than 99.5% of the segment boundaries were correctly identified and the variance among labelers was less than 10ms.[6] Also most of mis-labeled portions when checked by the rules were recovered. Mis-labeling rates will be

reduced by improvements in the labeling support system.

Moreover, some labeling comment system is recommended for labelers when any peculiar portion or an undefined phenomenon is found.

5 Conclusion

A large sized Japanese speech database at *ATR (JSDB-ATR)* was introduced. In particular, acoustic-phonetic transcriptions were precisely explained using some speech samples with their spectrograms. These multiple transcriptions are proposed for intelligent access to speech data and for fine acoustic-phonetic analysis. It is expected that these speech data can be quickly accessed through the speech database management system which is now under construction and that the JSDB-ATR will be effective for use in speech research.

Acknowledgments

We grateful to Dr. Shikano and Dr. Tohkura for their helpful discussions. We also wish to thank Dr. Kurematsu and Dr. Yodogawa for their continuous support of this work.

Reference

[1] Leung, H.C., and Zue, V. W., "A Procedure for Automatic Alignment of Phonetic Transcription with Continuous Speech," *Proc. ICASSP 84: IEEE International Conference on Acoustics, Speech, and Signal Processing*,1984, pp. 2.7.1-2.7.4.

[2] Carlson R. and Granstrom B., "A Search for Durational Rules in a Real-Speach Data Base," *Phonetica 43*, 1986, pp. 140-154.

[3] Itahashi S.,"Speech database of discrete words," *Journal of Acoustic Society Japan*, Oct. 1985, pp. 723-726.

[4] Tanaka K., Hayamizu S. and Ohta K., "Automatic labeling of known speech using a demiphoneme network representation and a parameter series segmentation technique," *Journal of Acoustic Society Japan*, Nov. 1986, pp. 860-868.

[5] Takeda K., Sagisaka Y. and Katagiri S., "Acoustic-Phonetic Labeling in a Japanese Speech Database," *Preprints Spring Meeting Acoustic Society Japan*, Mar. 1987, Paper 2-5-10.

[6] Katagiri S., Takeda K. and Sagisaka Y., "Characteristics of Phonetic Transcription Created by Spectrogram Observation," *Preprints Spring Meeting Acoustic Society Japan*, Mar. 1987, Paper 2-5-9.

Appendix

Hepburn system of romaji and supplementary expressions of non-native sounds

a	i	u	e	o
ka	ki	ku	ke	ko
sa	shi(si)	su	se(she)	so
ta	chi(ti)	tsu	te	to
na	ni	nu	ne	no
ha	hi	fu	he	ho
(fa)	(fi)		(fe)	(fo)
ma	mi	mu	me	mo
ya		yu		yo
ra	ri	ru	re	ro
wa				
n(transcribed by N)				
ga	gi	gu	ge	go
za	ji(zi)	zu	ze	zo
da	ji(di)	zu	de	do
ba	bi(vi)	bu	be	bo
pa	pi	pu	pe	po
kya		kyu		kyo
sha		shu		sho
cha		chu	(che)	cho
nya		nyu		nyo
hya		hyu		hyo
mya		myu		myo
rya		ryu		ryo
gya		gyu		gyo
ja		ju	(je)	jo
bya		byu		byo
pya		pyu		pyo

(non-native sounds)