TR－A－0039

Relaxation-based speech labeling

片桐　滋

Shgeru Katagiri

1988. 11. 24

ATR視聴覚機構研究所

U.7 Relaxation-based speech labeling

Shigeru Katagiri

ATR Auditory and Visual Perception Research Laboratories (MID Tower Bldg., Twin 21, 2-1-61 Shiromi, Higashi-ku, Osaka, Japan 540)

Abstract: It was revealed that a trained individual, i.e., a labeler, could perform accurate speech labeling and that such accuracy was based on his/her flexible decision process using many kinds of spectrographic features. In this paper, a new relaxation-based speech labeling system which duplicates the ability of the labelers is proposed. To realize the trial-and-error process of the labelers in the system, we have adopted a blackboard model and a discrete relaxation process.The system consists of a blackboard, and three subsystems: an acoustic analyzer, a verifier, and a supervisor. The blackboard is a working memory through which the three subsystems communicate with each other, and allows the system to realize the complicated behavior of the trial-and-error process. The acoustic analyzer computes many kinds of acoustic parameters, e.g., formant and pitch frequencies, corresponding to the spectrographic features used by the labelers. Also, the verifier is broken down into a symbol hypothesizer, and two kinds of functions: boundary detectors, and label identifiers. The verifier, with a behavior principle based on the relaxation process, efficiently performs the hypothesis verification for many of the label candidates. The supervisor controls the whole system. Preliminary experiment results show that the performance of the system is comparable to the performance of the labelers.

1. Introduction

A large-scale speech database with precise labels is indispensable to research on speech; here the label means a speech wave segment and its corresponding symbol which describes the acoustic characteristics within the segment. To construct such a database, one must overcome the inevitable difficult task of labeling. There are two possible approaches of labeling: a machine-labeling and a hand-labeling approach. Given any excellent labeling system, machine-labeling will obviously be more efficient than hand-labeling. However achieving the perfect system is unrealistic. We must therefore patiently carry out work corresponding to the hand-labeling and correct the machine-created labels. On the other hand, though time-consuming, hand-labeling by experts, i.e., labelers, will certainly produce the most desirable labels [Katagiri 88a]. Considering these aspects of the two approaches, we decided to adopt hand-labeling at least in building up the foundations of ATR speech databases [Takeda 87].

At present hand-labeling is proceeding smoothly, and the sets of speech waves and labels corresponding to several tens of thousands of words are available. We are now in the second step; the labeling task should be reconsidered from a new standpoint, where we have already acquired many kinds of heuristics on the labeling. We have accordingly started to design the labeling system. If the labeling system were available, it could reduce the time-consuming work of the labelers and further advance the standardization of label quality. Furthermore, study of the labeling system will contribute to research on speech recognition, because the main tasks in the labeling, i.e., speech segmentation and label identification, are also key tasks in recognition.

Promising approaches for the labeling or recognition system can be classified into a stochastic model-based approach and a knowledge-based approach. Recent works have suggested that the stochastic model-based system can be a viable candidate for a speech recognition system [Waibel 87][McDermott 88]. Accordingly, the stochastic model-based labeling system may also be hopeful. However, the stochastic model-based system requires enormous data to train itself, which is obviously a contradictory requirement, at least in the labeling work. We indeed need the labeling system to prepare the databases. On the other hand, the spectrographic features used by the labelers are clearly specified and their decision processes are explicit [ATR 88][Takeda 88]. As some speech recognition expert systems based on the spectrogram reading have been attempted [Zue 86], we can construct a labeling system duplicating the ability of the labelers. Such a knowledge-based system will consequently be one of pragmatic solutions to the machine-labeling problem.

In this paper, we present a new relaxation-based speech labeling system, which duplicates the ability of the labelers; the system is a kind of knowledge-based system. The hand-labeling is characterized by a flexible decision process, in other words, a trial-and-error process, using many kinds of spectrographic features [Katagiri 88a][ATR 88]. Accordingly, to duplicate this complicated process, we have adopted a blackboard model in the AI techniques [Hayes-Roth 85] and a discrete relaxation process [Rosenfeld 76][Katagiri 88b].The details and performance of this system will be described in the following paragraphs.

2. System Description

2.1 Overview

Our experiments have revealed that trained labelers are able to create accurate and consistent labels on a spectrogram, and also suggested that the labelers make their decisions in a very flexible trial-and-error process [Katagiri 88a]. Furthermore, in the daily work of hand-labeling, we can easily observe the following points; though the labelers don't quantitatively measure the spectrographic features, they skillfully accomplish difficult tasks: topological analysis of the spectrographic features and speaker adaptation in the labeling, etc. We think that their skill is mostly due to a flexible trial-and-error process.

To duplicate these advantageous aspects of hand-labeling, we have adopted the system structure shown in figure 1. The system is broken down into a blackboard and three subsytems: an acoustic analyzer, a verifier, and a supervisor. The system is based on a kind of a blackboard model in the AI area, which structure allows the system to realize a flexible decision processes. The three subsystems communicate with each other through this blackboard. The acoustic analyzer computes the acoustic parameters corresponding to the spectrographic features used by the labelers. Also, the verifier is broken down into symbol hypothesizer, and two kinds of functions: boundary detectors, and label identifiers.

The verifier, with a behavior principle based on the relaxation process, efficiently performs the hypothesis-verification for many label candidates. The supervisor controls the whole system.

The notations and terms used in this paper are shown in table 1. In the system, many terms and notations are defined; they will help us precisely understand the step-by-step behavior of the system. The details of the system are described in the following sections.

2.2 Blackboard

The idea of the blackboard is based on the blackboard model in AT systems. Using this idea, we can perform elaborate decision processes with different kinds of knowledge. In this labeling system, this blackboard plays a key role as a working memory through which the three subsystems communicate with each other; the subsystems and the functions included within the verifier pick up some entries from the blackboard as their inputs and enter their outputs onto the blackboard. All the parameters and candidates for the labeling shown table 1 are entered there.

2.3 Acoustic Analyzer

2.3.1 Acoustic Parameters

The acoustic parameters corresponding to the spectrographic features in the hand-labeling are computed in the acoustic analyzer. The pitch frequency [_P_] is derived from the cepstrum analysis. The formant frequencies [_Fx_] and bandwidths [_FBx_] (x=1,2) are computed with the root finding of the LPC model; F1 and F2 mean the first and second formant, respectively. A short-term power [_PW_], short-term band limited powers [_BPx_] (x=1,2,...,16) are also calculated; the frequency scaling in [_BPx_] is based on the Melscale spectra [Waible 81]. The spectrum change parameter is based on the fluctuation in LPC cepstrum coefficients [Sagayama 79]. The detailed specification of the acoustic analyzer is shown in table 2. Also, an output example of the acoustic analyzer is shown in figure 2; the labels by the hand-labeling are overwritten.

2.3.2 Estimation of Pitch and Formants

As shown in figure 2, the computed frequencies of the pitch and the formant, i.e., [_P_] and [_Fx_], are not desirable estimates. For example, we can find the pitch and formant contours during the label [cl] (closure), and also find discontinuity of the pitch contour during the labels, [i] or [o] (vowels). These parameters are particularly important for voiced/unvoiced decision and vowel categorization. To make a reliable decision, more precise estimates for the pitch and formants are obviously needed.

The precise estimation of the pitch and the formant requires several kinds of knowledge about acoustic phonetics; it is difficult to acquire the right estimates using only simple and straightforward signal processing algorithms. Therefore, the confidence scores for these estimates are calculated by using the knowledge base and other acoustic parameters. The knowledge base is prepared in the acoustic analyzer. We have named this procedure the knowledge-based confidence scoring

The principal idea to calculate the confidence scores is described here. We have utilized some requirements for the desirable characteristics of the pitch and the formants. These requirements are relative to the formant frequency [_Fx_], the formant bandwidth [_FBx_], the pitch frequency [_P_], short-term powers [_PW_] and [_BPx_], periodicity of the pitch (cepstrum coefficient value of the narrow peak which is expected to correspond to the right pitch frequency) [_C_], and continuity of the pitch and formant contours. Also two continuity parameters, [Pcont] and [Fcont], are defined for representing continuity of the pitch and the formant contours, respectively. Furthermore the bounded monotone functions which map the acoustic parameters to the confidence scores are designed according to the acoustic phonetics; here, the more reasonable the parameters, the higher the mapped confidence scores. The confidence scores are calculated at each discrete time index according to the following expressions;

Cp(n) = f1([_BPx_](n)) x f2([_C_](n)) x f3([Pcont](n)),
Cfx(n) = f4([_BPx_](n)) x f5([_FBx_](n)) x f6([Fcont](n)).

Cp(n): the confidence score for the pitch
Cfx(n): the confidence score for the x-th formant (x=1,2)
Here (n) represents the discrete time index. A function fj(y) (j=1,2,...,6) is a kind of bounded monotone function. The functions in addition to f5 are designed so that the larger their inputs, the larger their outputs. On the other hand, f5 is designed so that the narrower the [_FBx_], the larger the output of f5.

Also, to acquire a smoothed contours, we have adopted a median smoothing technique. In this smoothing, only the estimates with high confidence scores are used. The estimates with low confidence scores don't affect the resulting estimates.

2.4.Verifier

2.4.1 Overview

The verifier includes a symbol hypothesizer and two kinds of functions: boundary detectors and label identifiers. The task of the verifier is to perform the labeling under the relaxation process; this subsystem plays a central role in the whole system. The detailed behavior is described here.

As described before, the label consists of the speech segment and its corresponding symbol; in this system, 31 label symbols shown in table 3 are used. These are mostly the same as the symbols used in hand-labeling at ATR [Katagiri 88a][Takeda 87]. Furthermore, 463 pairs of label symbols are allowed to represent the adjacent labels. These limited pairs are selected out of all the label symbol combinations (31 x 31 combinations); the limitation is based on the labeling rules [ATR 88][Takeda 87].

As shown later, label identification is preceded by the boundary detection. The reasons for this are as follows.(1) Since the change of the acoustic parameters around the segment boundaries are dominant and consistent, the boundaries are easier to find than the centers of the segments.(2) If some function must cope with undesirable variations of the acoustic features due to co-articulation, the implementation of the function would be complicated and decrease the modularity of the system design; such a function would be complicated, with many rules. The simple boundary detector which is composed of a few simple rules and focused on the acoustic variations peculiar to every pair of labels should be positively implemented.

Accordingly, the detailed acoustic characteristics peculiar to the local region sandwiched between adjacent label segments are utilized to detect the boundary, and the gross but stable characteristics which can certainly appear in the whole label segment are used to identify the label category. These functions are designed according to the heuristics which we have already acquired in the hand-labeling work [Takeda 88][ATR 88].

2.4.2 Symbol Hypothesizer

The symbol hypothesizer translates a given speaking text, such as (i k i o i), into a string of possible labels, ([_i_][_>_][_cl_][_k_][_i_][_o_]); here the term "possible" means the possibility of appearing in the resultant labels. The label symbols which are not selected here will not appear in the outputs of the system

In this step, the location and duration of the possible label are not specified; they will be decided in later steps. All the outputs with a speaking text (i k i o i) given are shown in figure 3. In this figure, we can find surprisingly many variations of the strings with a short speaking text .

2.4.3 Boundary Detector

One boundary detector is prepared for every pair of adjacent label symbols. The boundary detector corresponding to the pair of adjacent possible labels is selectively triggered. Suppose that there are two possible labels [_x_] and [_y_] seen in the outputs from the symbol hypothesizer. Then the detector <BD-x-y> is selected and triggered. <BD-x-y> searches for the boundary candidate [_bd-x-y_] on a speech wave from left to right; a number of [_bd-x-y_] candidates will appear.

Now there are several boundary candidates on the blackboard. Selecting two of these candidates, we can set a segment. This process is described in detail here. Suppose that there are three boundary candidates: [_bd-i->_], [_bd-i-o_], and [_bd-o-i_]. If combining [_bd-i->_] with [_bd-i-o_], the segment sandwiched between these two boundary candidates would have incompatible label symbols, i.e., {>} and {i}. On the other hand, by combining [_bd-i-o_] with [_bd-o-i_], the segment between them would have a compatible label symbol, namely, {o}. The segment with the compatible label symbol is defined as the possible label [_o_]. Obviously the order in combining the boundary candidates is important; if and only if we combine the preceding boundary candidate with the following boundary candidate along with the given order, we will have a meaningful segment, i.e., a segment with positive duration. It is easy to understand that the process of searching for compatible label symbols, in other words, the possible labels, is one of discrete relaxation; according to the compatibility, many combinations of the boundary candidates are effectively reduced to a small set of the possible labels. The criterion in relaxation, i.e., compatibility, can be summarized as follows.

There are two boundary candidates: the preceding one, such as [_bd-x1-y1_], and the following one, such as [_bd-x2-y2_]. If the label symbol {y1} is the same as the label symbol {x2}, the segment sandwiched by these two boundary candidates is set as the possible label [_y1_], (y1=x2).

2.4.4 Label Identifier

It should be remembered that the boundary detector is focused on the local region around the boundary candidates rather than on the whole segment. We need to investigate the acoustic characteristics of the whole segment , and examine whether the possible label actually exists or not.

Here one label identifier is prepared for every label symbols. In this step, the label identifiers corresponding to the possible labels are selectively triggered; e.g., <LV-x> is triggered and examines whether the possible label [_x_] actually exists or not at the region where it was hypothesized beforehand, and emits a confidence score for [_x_]. Here, since the label identifier is triggered at the hypothesized region, it does not work from left to right.

2.5.Supervisor

The supervisor controls behavior of the whole system. This subsystem always monitors the blackboard and selectively triggers the subsystems and the functions in the verifier. After all the procedures are performed, the supervisor emits the possible labels with scores higher than the thresholds; these labels are the results of the system.

3. Experiment

3.1 Overview

We are now expanding many of the boundary detectors and label identifiers. The system is still not able to emit all kinds of labels. Thus, in this paper we have monitored system behavior and made a preliminary evaluation of the system performance though limited experiments. Speech data used here are the phonetically balanced 215 Japanese words uttered by two female speakers; they are parts of the data used in the previous experiments [Katagiri 88a].

3.2 System behavior

Looking at the blackboard, we can follow system behavior. Part of the blackboard with a speech wave and its corresponding speaking text (i k i o i) given to the system is shown in figure 4: the outputs from the boundary detectors in (4-a), and the outputs from the label identifiers in (4-b). We can see that only a few functions in the verifier are actually triggered under the relaxation.

3.3 Preliminary Evaluation of System Performance (V/UV Segmentation)

If an acceptable pitch estimate is found, the speech is characterized as voiced with the indicated pitch frequency; and if no reasonable pitch estimate can be found, the segment is identified as an unvoiced segment. Therefore, it is important to avoid spurious response in the unvoiced segment as well as to get accurate pitch frequencies. Similarly the lower formants are the key features in the voiced/unvoiced (V/UV) decision. Furthermore the formant frequencies contain important information for vowel categorization. The pitch and formant estimation are obviously quite relative to the V/UV decision and are important bases for the later decisions, e.g., the precise labeling.

In this experiment, to estimate the principal capability of the system, we focused on the performance of the V/UV segmentation, and actually evaluated the performance of the knowledge-based confidence scoring for the pitch and formants, and the V/UV segmentation. The definition of voicing is as follows: the voiced segment possesses the estimated contours of both the pitch and the 1st formant.

It is rather difficult to evaluate the accuracy of the estimated pitch and formant frequencies for natural speech waves; the reason for this is that the true values are unknown. Therefore we indirectly evaluated that scoring through V/UV decision performance. Here, the high performance in the V/UV decision was expected to guarantee accurate estimates of pitch and formants.

The estimated pitch and formant contours through knowledge-based confidence scoring and median smoothing are shown in figure 5. The acceptable pitch and formants are shown in this figure.

Also, the correct ratios of the V/UV decisions (%) are shown in figure 6. The ratio means the durational ratio of the region, categorized as the voiced segment, to the whole label segment. In figure 6, each bar graph indicates the average ratio for all the segments by two speakers, label by label. For example, when the region of the segment [p] is identified as the unvoiced segment, the decision is correct; when the region of the segment [a] is identified as the unvoiced segment, it is incorrect. Therefore, for the voiced labels, the higher the ratios, the more desirable the results; for the unvoiced labels, the lower the ratios, the less desirable the results. This figure shows that, although the system is still in the preliminary stage, the performance is mostly good. However, we obviously need to improve the functions in the verifier; in particular, the functions related to the liquid ([r]) and the plosives ([p], [t], [k], [b], [d], [g]) should be sharply improved. These label segments with low performance are of short duration. The analysis conditions in the acoustic analyzer would be somewhat inappropriate. To resolve this difficulty, we have to add some rules to compensate the acoustic analyzer outputs.

4. Conclusion

In this paper, we proposed a relaxation-based speech labeling system, and showed the preliminary experiment results; we showed the principal behavior of the system, and we evaluated the performance focused on the V/UV segmentation. It was consequently revealed that, although the implementation of the system was still in a preliminary stage, the system accomplished good segmentation. Encouraged by this result, we are now continuing to extend the system.

Acknowledgement

This research is closely related to the Japanese speech data base construction by ATR Interpreting Telephony Research Laboratories and ATR Auditory and Visual Perception Research Laboratories. The knowledge acquisition for function implementation in the system is supported by the labelers and the members of the spectrogram reading training course at ATR. The author would like to thank all of these people.

References

[ATR 88] ATR Interpreting Telephony Research Laboratories; "Principal procedures for speech labeling (version 8)", ATR internal report, June, 1988 (in Japanese).
[Katagiri 88a] S.Katagiri, K.Takeda, Y.Sagisaka; "Speech labeling Using a Spectrogram", IEICE, Speech Study Group Meeting Report SP87-115, Vol.87, No.350, pp.15-22, January, 1988 (in Japanese).
[Katagiri 88b] S.Katagiri; "Knowledge Based Acoustic Analysis of Speech Signal", ASJ, Fall Conference, 2-1-19, Vol.1,pp.203-204, March, 1988 (in Japanese).
[McDermott 88] E.McDermott and S.Katagiri; "Shift-tolerant, Multi-phoneme Recognition Using Learning Vector Quantization", Speech Study Group Meeting Report, October, 1988.]
[Hayes-Roth 85] F.Hayes-Roth, D.Waterman, and D.Lenat; "Building Expert Systems (Japanese edition)", pp.133-137, 1985.
[Resenfeld 76] A.Rosenfeld, R.Hummel and S.Zucker; "Scene Labeling by Relaxation Operations", IEEE Trans. SMC, Vol.SMC-6, No.6, pp.420-433, June, 1976.
[Sagayama 79] S.Sagayama and F.Itakura; "On Individuality in a Dynamic Measure of Speech", ASJ, Spring Conference, Vol.2, 3-2-7, pp.589-590, June, 1979 (in Japanese).
[Takeda 87] K.Takeda, Y.Sagisaka, and S.Katagiri; "Acoustic-Phonetic Labels in a Japanese Speech Database", Proc. European Conference on Speech Technology. Vol.2, pp.13-16, 1987.
[Takeda 88] K.Takeda, Y.Sagisaka, S.Katagiri, and H.Kuwabara; "Manual Segmentation of Spectrogram for the Acoustic-phonetic Transcriptions in a Japanese Speech Database", TR-I-0019 (TR-A-0019), ATR, February, 1988 (in Japanese).
[Waibel 81] A.Waibel and B.Yagnanarayana; "Comparative Study of Nonlinear Time Warping Techniques in Isolated Word Speech Recognition Systems", CMU-CS-81-125, Carnegie-Mellon University, June, 1981.
[Waibel 87] A.Waibel; "Phoneme Recognition Using Time-delay Neural Networks", IEICE, Speech Study Group Meeting Report, Vol.87, No.299, pp.19-24, SP87-100, 1987.
[Zue 86] V.Zue and L.Lamel;" An Expert Spectrogram Reader: A Knowledge-based Approach to Speech Recognition", IEEE, ICASSP 86, Vol.2, 23.2, pp.1197-1200, April, 1986.

Table 1 System notations and terms.

- PART I - The following are used for general discussions.	
{x}	label symbol x
{x-y}	pair of adjacent label symbols, {x} and {y}
[x]	label associated with the label symbol {x}
[bd-x-y]	boundary between the adjacent labels, [x] and [y]
- PART II - The following are notations or terms used in the system.	
acoustic feature	pitch frequency [P] (Hz)
	formant frequency [Fx] (Hz) (x=1,2)
acoustic parameter	pitch frequency [_P_] (Hz)
	cepstrum amplitude corresponding to pitch peak [_C_]
	formant frequency [_Fx_] (Hz) (x=1,2)
	formant bandwidth [_FBx_] (Hx) (x=1,2)
	short-term power [_PW_] (dB)
	band-limited short-term power [_BPx_] (dB) (x=1,2,...,16)
	spectrum change parmeter [_SC_]
[_x_]	label candidate for [x] (possible label)
[_bd-x-y_]	boundary candidate for [bd-x-y]
<LV-x>	label identifier for [x]
<BD-x-y>	boundary detector for [bd-x-y]

Table 3 Symbols for labels.

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

Table 2 Acoustic analyzer specification.

pre-emphasis	first-differencing
time window	Hamming window of 30 msec shift interval of 2.5 msec
FFT	512 points
LPC	13 poles autocorrelation method
LPC cepstrum	13 coefficients
pitch	FFT cepstrum based detection
Mel spectrum	16 coefficients [Waibel 81]

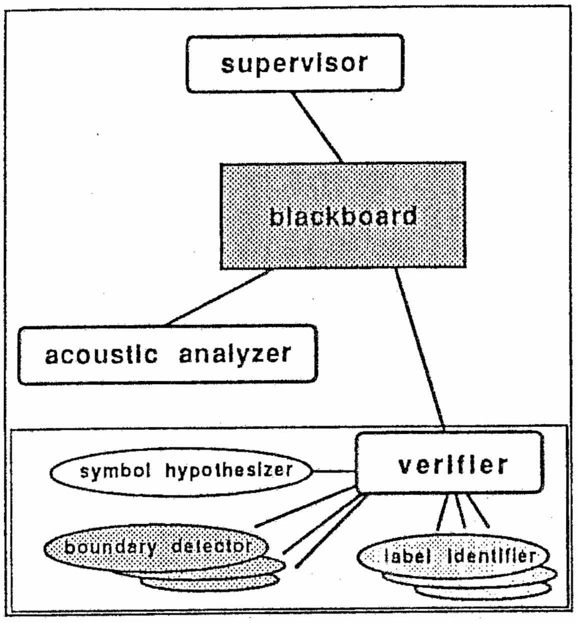

Figure 1 Structure of relaxation-based speech labeling sytem.

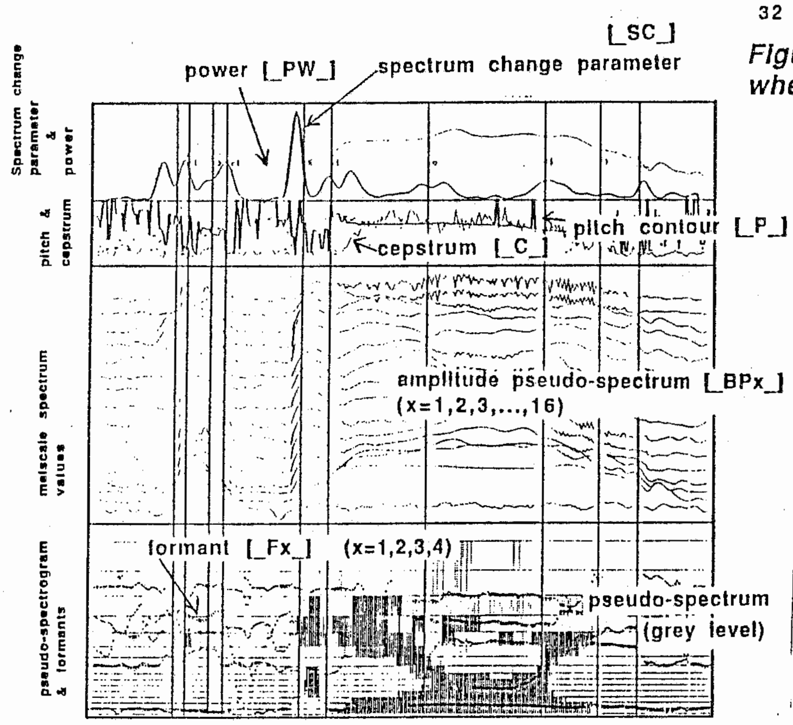

Figure 2 An example of outputs from the acoustic analyzer.

The figure is divided into four parts. Contours for a spectrum change parameter and power are shown in the first part; pitch and cepstrum contours in the second part, 16 melscale spectrum values in the third part, and formants and density pseudo-spectrogram in the fourth part. A dark region on this spectrogram is an energy concentrated region. Moreover, a dark circle in the lowest part shows a formant with a narrow bandwidth. The 18 horizontal lines in the pseudo-spectrogram illustrate distribution of melscale channels.

The notations used here are shown in Table 1.

```
(i k i o i) ──┐
              ▼
 1  ([_pau_][_<_][_i_][_>_][_cl_][_k_][_i_][_o_][_i_][_>_][_pau_])
 2  ([_pau_][_i_][_>_][_cl_][_k_][_i_][_o_][_i_][_>_][_pau_])
 3  ([_pau_][_<_][_i_][_cl_][_k_][_i_][_o_][_i_][_>_][_pau_])
 4  ([_pau_][_<_][_i_][_>_][_cl_][_k_][_o_][_i_][_>_][_pau_])
 5  ([_pau_][_<_][_i_][_>_][_cl_][_i_][_o_][_i_][_>_][_pau_])
 6  ([_pau_][_<_][_i_][_>_][_cl_][_k_][_i_][_o_][_i_][_pau_])
 7  ([_pau_][_i_][_cl_][_k_][_i_][_o_][_i_][_>_][_pau_])
 8  ([_pau_][_i_][_>_][_cl_][_k_][_o_][_i_][_>_][_pau_])
 9  ([_pau_][_i_][_>_][_cl_][_i_][_o_][_i_][_>_][_pau_])
10  ([_pau_][_i_][_>_][_cl_][_k_][_i_][_o_][_i_][_pau_])
11  ([_pau_][_<_][_i_][_cl_][_k_][_o_][_i_][_>_][_pau_])
12  ([_pau_][_<_][_i_][_cl_][_i_][_o_][_i_][_>_][_pau_])
13  ([_pau_][_<_][_i_][_cl_][_k_][_i_][_o_][_i_][_pau_])
14  ([_pau_][_<_][_i_][_>_][_cl_][_o_][_i_][_>_][_pau_])
15  ([_pau_][_<_][_i_][_>_][_cl_][_k_][_o_][_i_][_pau_])
16  ([_pau_][_<_][_i_][_>_][_cl_][_i_][_o_][_i_][_pau_])
17  ([_pau_][_i_][_cl_][_k_][_o_][_i_][_>_][_pau_])
18  ([_pau_][_i_][_cl_][_i_][_o_][_i_][_>_][_pau_])
19  ([_pau_][_i_][_cl_][_k_][_i_][_o_][_i_][_pau_])
20  ([_pau_][_i_][_>_][_cl_][_o_][_i_][_>_][_pau_])
21  ([_pau_][_i_][_>_][_cl_][_k_][_o_][_i_][_pau_])
22  ([_pau_][_i_][_>_][_cl_][_i_][_o_][_i_][_pau_])
23  ([_pau_][_<_][_i_][_cl_][_o_][_i_][_>_][_pau_])
24  ([_pau_][_<_][_i_][_cl_][_k_][_o_][_i_][_pau_])
25  ([_pau_][_<_][_i_][_cl_][_i_][_o_][_i_][_pau_])
26  ([_pau_][_<_][_i_][_>_][_cl_][_o_][_i_][_pau_])
27  ([_pau_][_i_][_cl_][_o_][_i_][_>_][_pau_])
28  ([_pau_][_i_][_cl_][_k_][_o_][_i_][_pau_])
29  ([_pau_][_i_][_cl_][_i_][_o_][_i_][_pau_])
30  ([_pau_][_i_][_>_][_cl_][_o_][_i_][_pau_])
31  ([_pau_][_<_][_i_][_cl_][_o_][_i_][_pau_])
32  ([_pau_][_i_][_cl_][_o_][_i_][_pau_])
```

Figure 3 Outputs from symbol hypothesizer when a speaking text (i k i o i) is given.

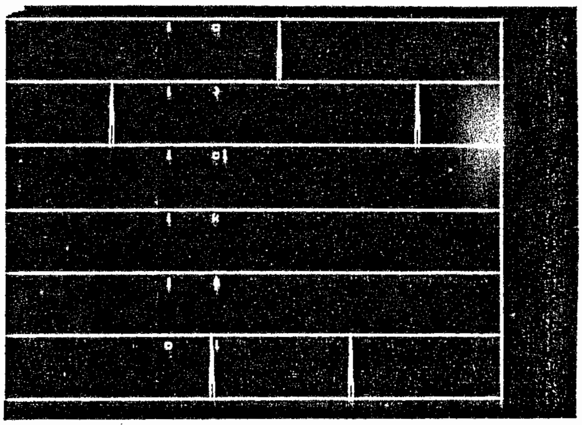

Figure 4 An example of entries on blackboard.

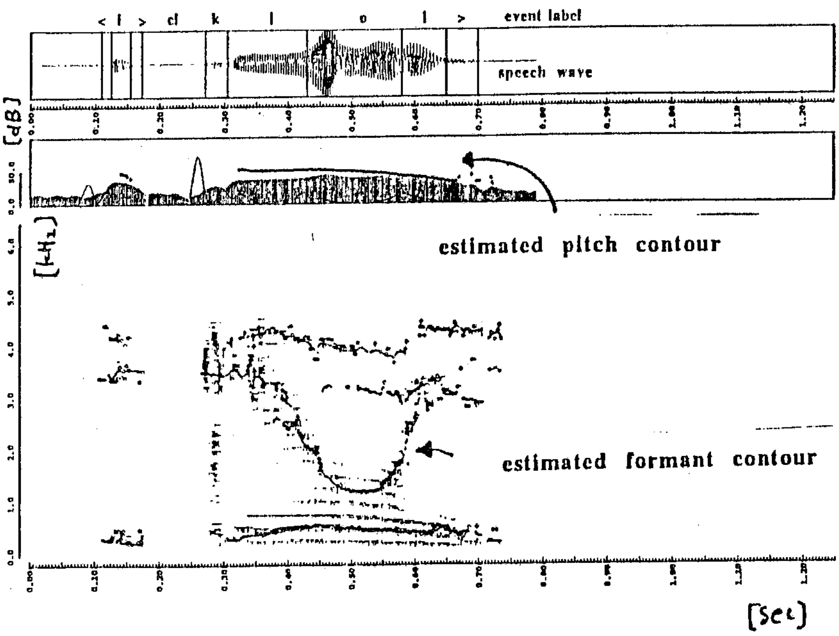

Figure 5 An example of estimated pitch and formant contours.

These contours are refined by knowledge-based confidence scoring and median-smoothing.

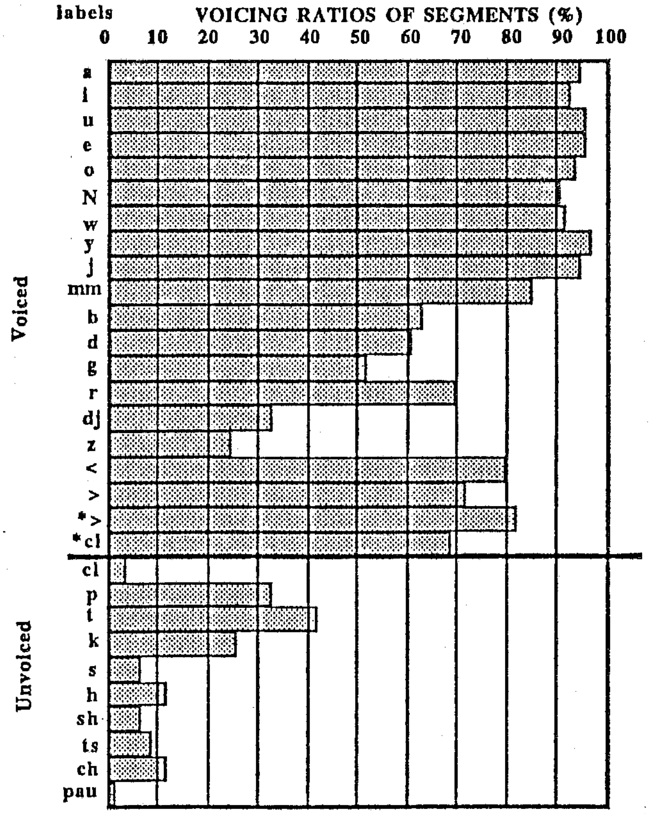

Figure 6 Performances in voiced/unvoiced decisions.

The voicing ratio means the ratio of the segment duration which was categorized as voiced to the whole label segment duration. Here, each bar graph indicates the ratios for all segments by two speakers, label by label. For the voiced labels, the higher the ratios, the more desirable the results; for the unvoiced labels, the lower the ratios, the more desirable the results.