非公開

TR-I-0081

An approach for normalizing coarticulatory variations through expansion of acoustic properties into phonetic features

音素的特徴の動的性質を用いた調音結合の正規化

Kazuya Takeda

武田一哉

1989.5

Abstract

This report proposes an approach for normalizing coarticulatory variations in connected speech. The proposed scheme expands acoustic realization of speech into distinctive features and on the basis of dynamic property of each feature, normalizes them independently. In this report, the effectiveness of the scheme is discussed focusing on an phonetic feature "voiced". First, for the extraction of the "voiced" feature from the speech, a static measurement is designed by optimizing the parameters of preliminary measurements based on acoustic-phonetic knowledges. Second, the extracted feature's dynamics are analyzed based on the discrimination performance of the static measurement. The experiment suggests that the dynamics of the measurement can be a good cue for detecting voice-onset and offset timing. Finally, a measurement which combines static and dynamic measurements shows 12% better accuracy than the static measurement. Through these experiments, the effectiveness of the proposed scheme in normalizing the coarticulatory variation is confirmed.

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

1. Introduction

The existence of coarticulatory variations of the acoustic realization of speech, especially fluently spoken speech, has been one of the serious problems of automatic recognition of the spoken languages. Various research works have been reported for the normalization of these variations.

Kuwabara reported an experiment, in which he tried to calculate the target formant frequencies of vowels which are affected by the formant locations of the adjacent vowels [1]. His work was motivated by the characteristic of human perception of vowels in connected speech, which highly depends on the context. In that experiment, he averaged the frequencies of the lowest two formants using a weighting function in the time domain, and succeeded in increasing separation of vowels in the $f_1 - f_2$ space.

Akagi *et al.* proposed a spectral target prediction method based on the 2nd-order critical damping model in [2]. They evaluated the prediction model as a preprocessor for a speech recognition system, and concluded its effectiveness.

Mainly, these approaches are based on the standpoint of simulating the compensation mechanism which presumably exists in speech perception mechanism of human auditory system [3]. On the other hand, from the standpoint of speech production, coarticulation can be regarded as a result of dynamic characteristics of articulators. Thus, if the speech signal can be expanded into some phonetic features associated with articulation (e.g. "voiced" with glottal vibration), then coarticulatory variations may be described by dynamic characteristics of these features. In this report, with the aim of evaluating the above scheme of normalizing coarticulatory variations, some preliminary experiments are described using a phonetic feature "voiced" as an example.

In section 2, acoustic measurements which extract the phonetic feature "voiced" from a static representation of speech are discussed. In this section, a measurement optimizing method (SAILS) is introduced. The measurements obtained using this method were compared for performance of "voiced" feature discrimination.

The results of a "voiced" feature discrimination experiment using these measurements were analyzed in section 3. Through error analysis of the experiment, it is found that with only static information, discrimination accuracy is not satisfactory.

The dynamic characteristics of the feature "voiced" is analyzed in section 4. Through the analysis, it is found that the change of the feature in time at voice-onset and voice-offset points is statistically stable. Finally, a "voiced/unvoiced" discrimination experiment showed that combining this dynamic property and the static information, produced a 12% higher accuracy than the static measurement above.

2. Design of spectral measurements for feature extraction

To analyze the dynamic characteristics of a phonetic feature, it is necessary to extract its movement from the speech signal. The movement can be represented as a time sequence of values which express how much each speech frame is under the influence of a feature. Given this requirement, the first step of this study was to design measurements for static spectral representations which measure how likely each frame has the acoustic property of the phonetic feature.

2.1 SAILS

There have been many attempts to design such feature extractors that can classify segments into phonetic categories. In this study, to obtain measurements which extract phonetic features, the newly developed SAILS system was used [4]. SAILS is a system in which the determination of an acoustic measurement for phonetic classification is formulated as a constrained search problem using knowledge of acoustic-phonetics. The measurement determination procedure of SAILS can be summarized as follows.

Using acoustic-phonetic knowledge, a primitive measurement or their combination can be selected as a phonetic classifier. For example, conventional investigation implies that energy of a low frequency band may be a reasonable cue for discrimination of voiced and unvoiced segments. Since measurements may have some free parameters, (e.g. to calculate the energy in a frequency band, lower and higher bounds of the band need to be specified) designing a specific measurement becomes a problem of optimizing these parameters. The optimization can be accomplished using classification performance on the training data for a particular task.

2.2 Experiment to optimize measurements

The experiment of designing measurements was carried out by using seven sentences uttered by a female speaker in the TIMIT database [5]. The phonetic class for each segment was obtained from the hand transcription of the database. In the experiment, three primitive measures; "Spectral Moment", " Low Frequency Energy" and "Low Frequency Energy Ratio" were compared in their classification performance. Definitions for each measurement are as follows.

$$\text{Spectral Moment} = \int_{f_1}^{f_2} f E(f) df \tag{1}$$

$$\text{Low Frequency Energy} = \int_{f_1}^{f_2} E(f) df \tag{2}$$

$$\text{Low Frequency Energy Ratio} = \frac{\int_{f_1}^{f_2} E(f)df}{\int_0^{\frac{F}{2}} E(f)df} \tag{3}$$

Two different spectral representation were used, a spectrum obtained by averaging frames over the segment and a spectrum at the center of the segment. The result of the experiment on feature "voiced" is illustrated in Figure 1. In this figure, normalized distance between the mean values of the categories (eqn. 4) was used as the index of classification performance.

$$\text{classification index} = \frac{(\mu_A - \mu_B)^2}{\sigma_A^2 + \sigma_B^2} \tag{4}$$

The combinations of the free parameters' values which maximize the performance of each measurement are shown in **Table 1,2**. This experiment showed that the most powerful classifier was "Spectral Moment" calculated from 0 to 2.5 [kHz] at the center of segment. There was not much difference in the performance of classification between the two spectral representations of input, however, optimized parameter's values are slightly depend on the representation of input.

From this result, it is seen that "Spectral Moment" below approximately 2,500 Hz is a good measurement for discriminating voiced from unvoiced segments. The ratio of energy in low frequency bound to overall energy is obtained approximately the same classification performance as the "Spectral Moment." However, the frequency bound is considerably different.

Table 1: Optimized parameters for each measurement [Hz]
input is the spectrum at the center of the segment.

Measurement	Lower bound	Higher bound
Spectral Moment	0.00	2500.00
Low Frequency Energy	0.00	468.25
Low Frequency Energy Ratio	0.00	781.25

Table 2: Optimized parameters for each measurement [Hz]
input spectrum is the averaged spectrum over the segment.

Measurement	Lower bound	Higher bound
Spectral Moment	0.00	2656.00
Low Frequency Energy	156.00	468.25
Low Frequency Energy Ratio	0.00	1562.50

2.3 Frame base classification experiment

In the previous section, parameters of three measurements were optimized from the standpoint of separating the distributions of each class. Since the final goal of this study is to formulate the feature dynamics, the measurement is needed to extract the phonetic feature of the frames as well as the segments. Thus, frame based classification performance is another important performance index of measurements. To examine the frame classification performance of the measurements, an experimental discrimination of "voiced" feature of each frame was carried out. In the experiment, the following discrimination criterion was used for each frame.

$$\text{The frame is} \begin{cases} \text{"voiced"} & \text{output of measurement} > \text{discrimination threshold} \\ \text{"unvoiced"} & \text{otherwise} \end{cases} \tag{5}$$

In the above discrimination, phonetic class of the frame is identified based only on the static characteristics extracted by each measurement. As an example of the search for the optimal threshold value, the relationship between the classification performance and the threshold value is illustrated for the "Low Frequency Energy" measurement in **Figure 2.** After optimizing the threshold value, frame classification performance is illustrated for each measurement in **Figure 3.** It can be seen that the highest performance was obtained using the "Low Frequency Energy" measurement and that the performance of it is about 90%.

On the basis of these results, the "Low Frequency Energy" measurement will be used as the extractor of the phonetic feature "voiced", in further sections.

3. Feature identification error analysis

In the previous section, through the designing of measurements for extracting the feature "voiced", the classification performance based on the measurement is obtained to be about 90%. In this section, the dynamic characteristic of the feature "voiced" will be discussed by analyzing the discrimination error of the experiment in which each frame's phonetic feature is discriminated with only static property extracted by the "Low Frequency Energy" measurement. Since the dynamic characteristics of the measurement is not taken account in the experiment, some systematic errors are expected to clarify the dynamic characteristics of the feature.

3.1 Error analysis according to the segment type

The segments appearing in the experiment include examples of 60 out of 64 labels in TIMIT phonetic transcription, and discrimination errors were observed for 52 labels. In **Figure 4**, the segments in which frame discrimination error

exceeds 10% were illustrated. In this figure 84% of mis-identified frames are included. From the figure, it can be seen that most of errors, actually 64% of all, are observed in voiced fricatives and stops. This result implies that even by using the optimized parameters, it is difficult to identify the phonetic feature "voiced" of the frame in voiced fricatives and stops, and is also in agreement with the difficulty in discriminating voiced from unvoiced fricatives in phoneme recognition.

In the other labels, the frame identification rate is quite good except for the burst portion of "k" in which the "compact burst" was often mistaken for voicing.

3.2 Error analysis according to context

Generally, contextual effects are closely related to coarticulatory phenomena. Thus, the relationship between discrimination performance and context is expected to suggest important characteristics of the feature "voiced". To examine the contextual effect on these mis-identifications, the errors in the voiced segments, are plotted with the context in **Figure 5**. Comparing left and right contexts, the effect of the left context seems to have the greater effect on performance. The figure reveals that "voiced" frames can be identified more accurately if its segment is preceded by an unvoiced segment than by a voiced segment, and that the identification is more accurate when it is followed by a voiced segment than by an unvoiced segment.

From the result of the previous section, that most of mis-identifications of "voiced" frames are found in voiced fricatives and stops, these contextual effect can be interpreted as follows, 1) Since the main cue of such segments is prevoicing, if the segments are preceded by a unvoiced segment, "voiced" feature is emphasized in the segment, on the other hand 2) if preceded by voiced segment, "voiced" feature is not emphasized in the segment.

3.3 Error analysis according to segmental duration

The typical coarticulatory variations are observed when the speech portion is under an articulatory influence, such as nasalization of vowels in nasal context. Mis-identifications in voiceless segments are mainly interpreted to be due to this phenomena. The feature "voiced" sometime does not change suddenly at voice-onset and offset position, or the "Low Frequency Energy" measurement can not locate these changes accurately. Unlike the errors in the voiced segments which can be interpreted as due to the phonetic characteristics of the segment, the errors in voiceless segments were mainly due to this mis-location of the phonetic boundary. The relationship between identification performance and the segment's duration, illustrated in **Figure 6**, reveals the above characteristics. In the figure, the ratio of mis-identified frames in the segment is plotted as a function of the segmental duration for all voiceless segments which are between

two voiced segments. As shown in the figure, the duration of the segments and resulted error rates are inverse proportional with a regression coefficient of -0.48. Especially, in the two samples whose duration are less than 50 milliseconds, very high error rates were observed.

4. Dynamic characteristics of the feature "voiced"

The results of the error analysis implied that with only the static measurement, the "voiced" feature can not be identified sufficiently in voiced fricatives and stops, or in short voiceless segments. In **Figure 7**, the output of the "Low Frequency Energy" measurement and "voiced/unvoiced" discrimination result based on the output of the measurement are illustrated, in the third and bottom window, for the utterance "The angry boy answered but didn't look up." In that figure, the errors described in the previous section can be seen as follows.

- The voiced stops are not identified as a "voiced" segment (/d/ and /b/ sounds in "answered but").
- The voice onset and offset positions are not located accurately in voiceless segment sandwiched by voiced segments (/t/ sound in "didn't").

However, dynamic characteristics of the output of the measurement, such as inclination at the voice-onset and offset positions in the figure, look stable and seem to be a better cue for locating voice-onset and offset positions. In this section, the use of such dynamic characteristics of the feature "voiced" is examined for normalization of coarticulatory variations in connected speech.

4.1 Analysis on differential values of low frequency energy

Figure 7 suggested that the differential values are good cues for detecting voice-onset and offset positions. To examine the stability of the value at that positions, distributions of differential values were analyzed. In **Figure 8**, the distributions of "Low Frequency Energy" values and its differential values using a 15 millisecond window are illustrated. In the figure, the white circle and associated line represent the mean value and standard deviation, and the black circle and associated line represent the distribution of the values at the voice-onset position.

From the distribution of the "Low Frequency Energy" values, it can be seen that the distance between mean value and optimal discrimination boundary (-92.0dB) is only 5% of the standard deviation of all data. However, in the distribution of differential values, the distance between the mean values of all data and the data at voice-onset position is more than 300% of the standard deviation of all data. These results show that the differential values of the "Low Frequency Energy" are better cues than the static values, for the location of

"voiced/unvoiced" boundary.

4.2 Feature identification using dynamic characteristics

The results of the previous section suggests that by using the differential values of the "Low Frequency Energy" as dynamic characteristics of the feature, voice-onset and offset positions can be located more accurately. In this section, a "voiced" frame discriminating experiment is carried out based on an algorithm using this result. The algorithm decides onset and offset of the "voiced" portions using not only output value of the "Low Frequency Energy" measurement but also its differential values. The detection algorithm is as follows.

1. Based on the "Low Frequency Energy" and its differential value, detect stable portion of the "voiced" segments.
2. Search backward for the voice-onset position of the "voiced" segment from the detected stable portion using the differential values. In this experiment, 15 [dB/15ms] was used for the threshold value for the detection.
3. Search forward for the voice-offset position of the segment from the stable position using the differential values. In this experiment, -10 [dB/15ms] was used for the threshold value for the detection.

An example of the experimental result is illustrated in Figure 9. In the figure, some of the voiced stops which were mis-identified in the previous experiment are correctly identified, and compact burst of "k" is discriminated as not "voiced". However, the mis-identifications in the voiceless stops are still not recovered.

Based on this procedure, mis-identified frames in "voiced" feature discrimination was reduced 12% from the discrimination based only static measurement. Mainly, the improvement of the performance was observed in voiced fricatives, which static measurement was not able to identify as voiced segments. As an example, performance of the two measurements are described in the Table 3 for the segment "z." The ratio of the correctly identified frames is increased to 78% by the algorithm described above from the 22% using the static measurement. As shown in the table the improvement of the frame discrimination accuracy resulted in improvement also in the accuracy of the feature identification of the segment, especially in tasks such as the discriminating of "z" from "s" in which identification of the "voiced" feature is distinctive.

Table 3: Feature identification ratio
The segment feature was identified using the center frame of the segment.

	Static Measurement	Dynamic & Static Measurement
Frame	22%	78%
Segment	0%	75%

From this experiment, it can be concluded that by introducing knowledge on dynamic characteristics, identification accuracy of the feature "voiced" was improved about 12%. However, further study is needed to improve the misidentification of the "voiced" feature in short voiceless segments.

5. Conclusions

In this report, an approach to the normalization of coarticulatory variations in the acoustic realization of connected speech was discussed. The approach is based on the scheme that expands acoustic properties into phonetic features, and formulates dynamic characteristic of each feature independently. The effectiveness of the scheme was confirmed through a set of experiments and discussions. First, the extraction of a phonetic feature "voiced" was carried out using a static measurement which obtained by optimizing parameters of preliminary measurements based on acoustic-phonetic knowledge. Second, an experiment of feature identification based only on the static measurement was revealed that with only static information, phonetic feature "voiced" could not be identified accurately. The result of error analysis on the experimental identification suggested that most of the resulting errors are due to articulatory effects of their context, and by using dynamic characteristics of the feature, discrimination error can be decreased. Finally, a feature identification experiment using both information of the static and dynamic properties of the feature, resulted in a 12% decrease of error rate.

Acknowledgements

This experimental works were carried out while the author was studying at Massachusetts Institute of Technology as a visiting scientist, with much help from Dr. Zue and his crew to whom the author is grateful. The author is also grateful to Dr. Kurematsu for encouragement and continuous support of the research and to Dr. Shikano and Mr. Phillips (Spoken Language Systems Group) for discussion and suggestive comments.

References

[1] "An approach to normalization of coarticulation effects for vowels in connected speech," Hisao Kuwabara, J.Acoust.Soc.Am. 77 (2), 1985.
[2] "On the application of spectrum target prediction model to speech recognition," Masato Akagi and Yo'ich Tohkura, proc ICASSP 1988.
[3] "On the role of formant transitions in vowel recognition," Lindblom, B.E.F., and Studdert-Kenedy, M., J.Acoust.Soc.Am. 58, 923-927, 1967
[4] "Automatic discovery of acoustic measurements for phonetic classification," Michael S. Phillips, the second Joint meeting of Acoust.Soc. of Am. and Jpn., November 1988.

[5] "Transcription and alignment of the TIMIT database," Stephanie Seneff and Victor W. Zue, *to be distributed with TIMIT database*, NBS, 1988

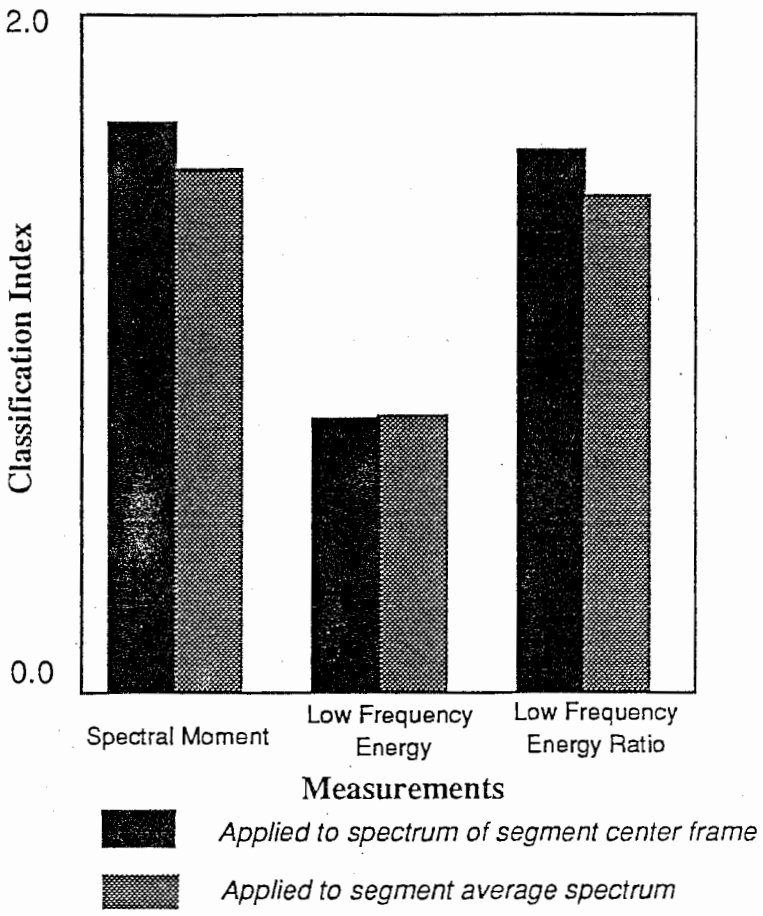

Figure 1: Classification performance using the classification index defined in eqn. 4

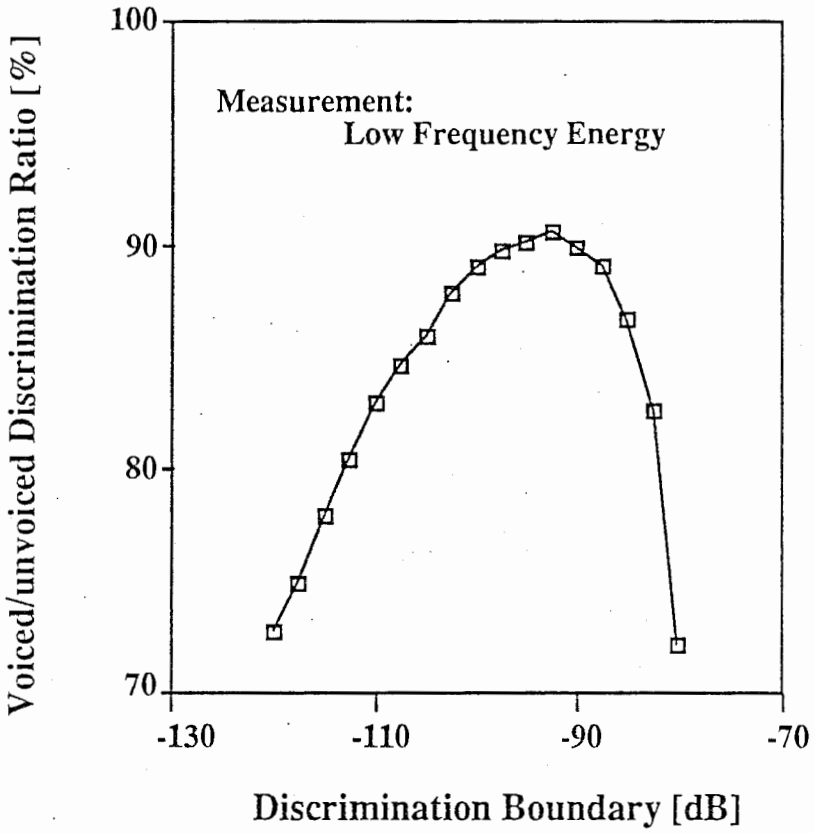

Figure 2: Optimal discriminating threshold

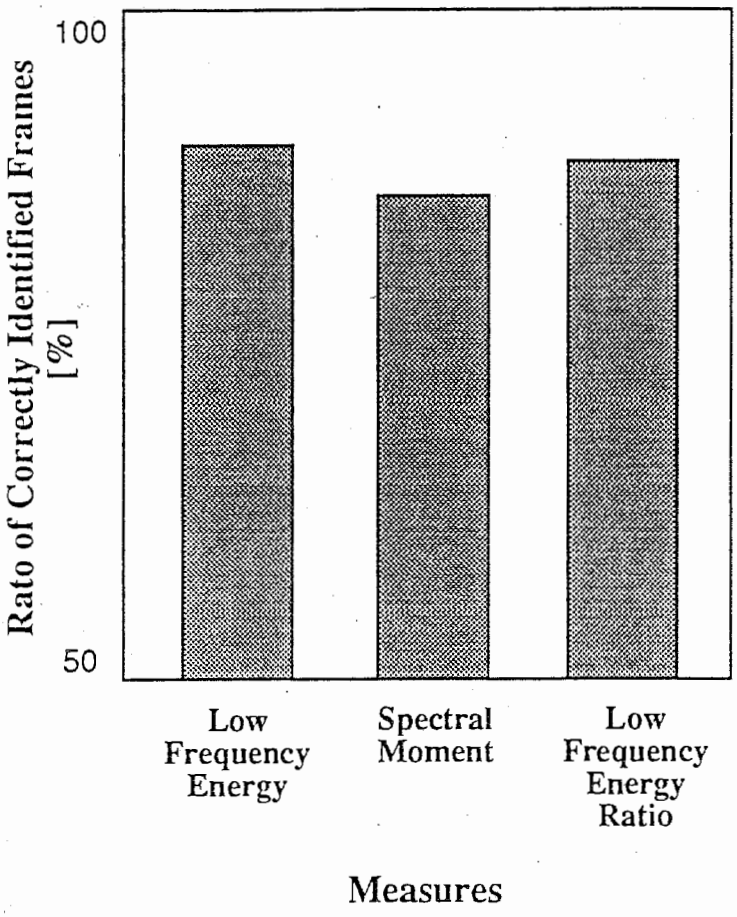

Figure 3: Classification performance of measurements evaluated by the frame classification rate

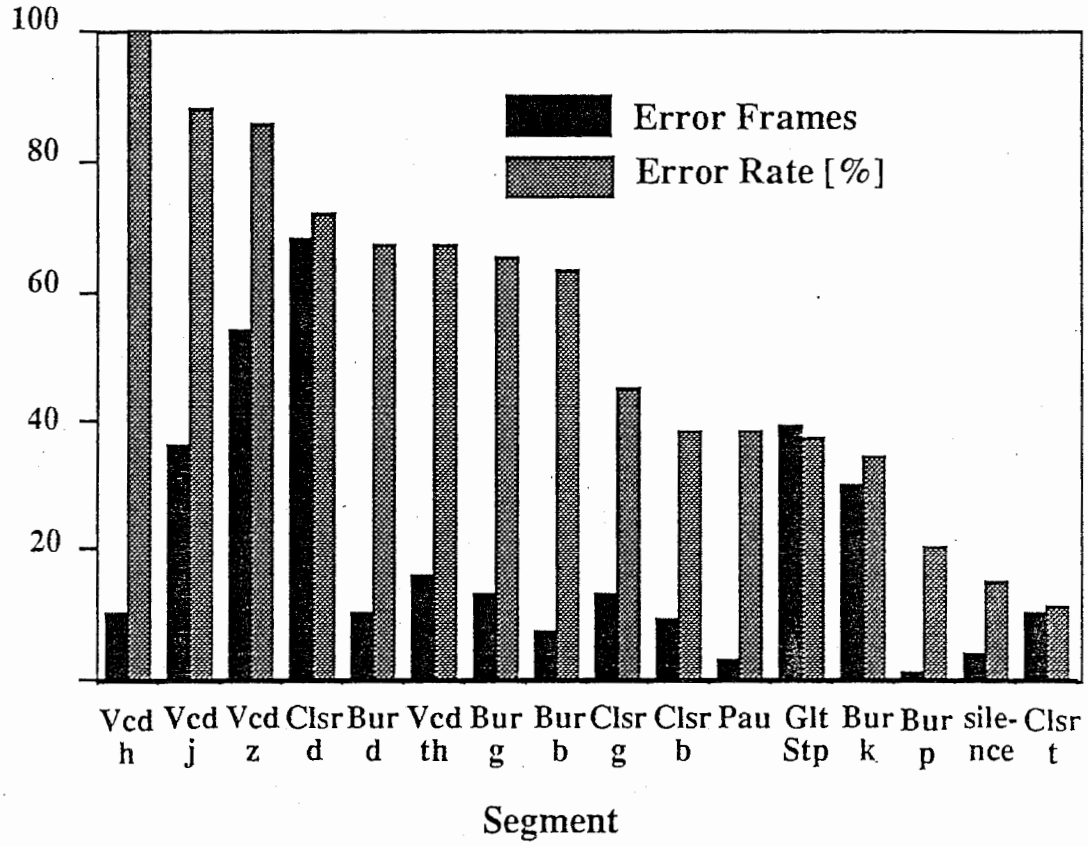

Figure 4: Mis-identified frames in each segment

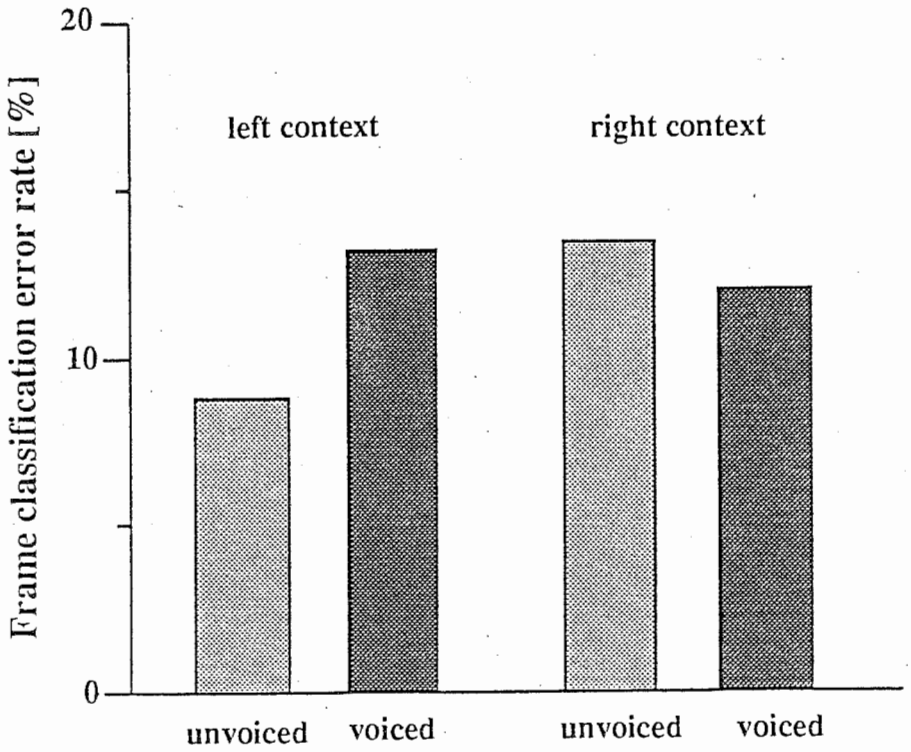

Figure 5: Mis-identified frames in each context of the segment

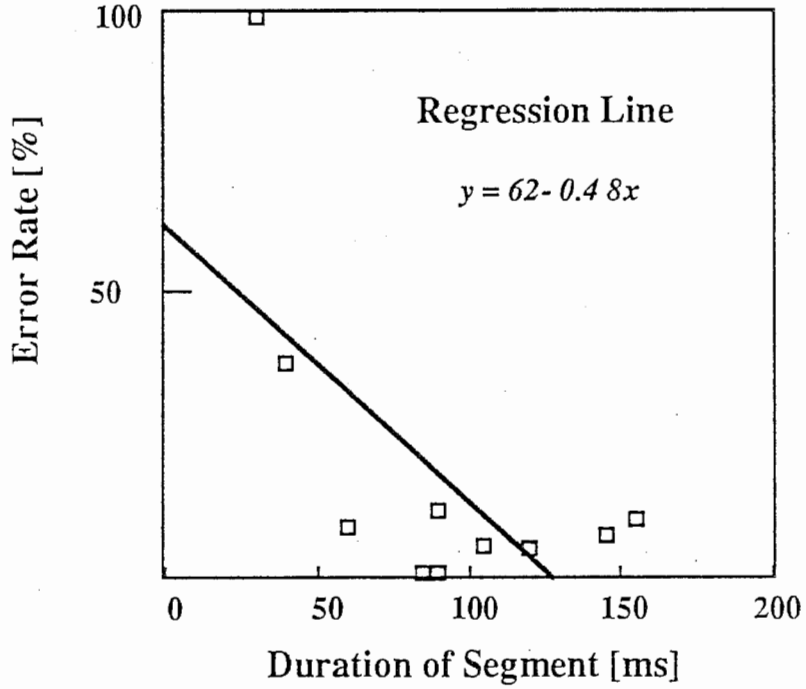

Figure 6: Mis-identifying rate and the duration of the segment

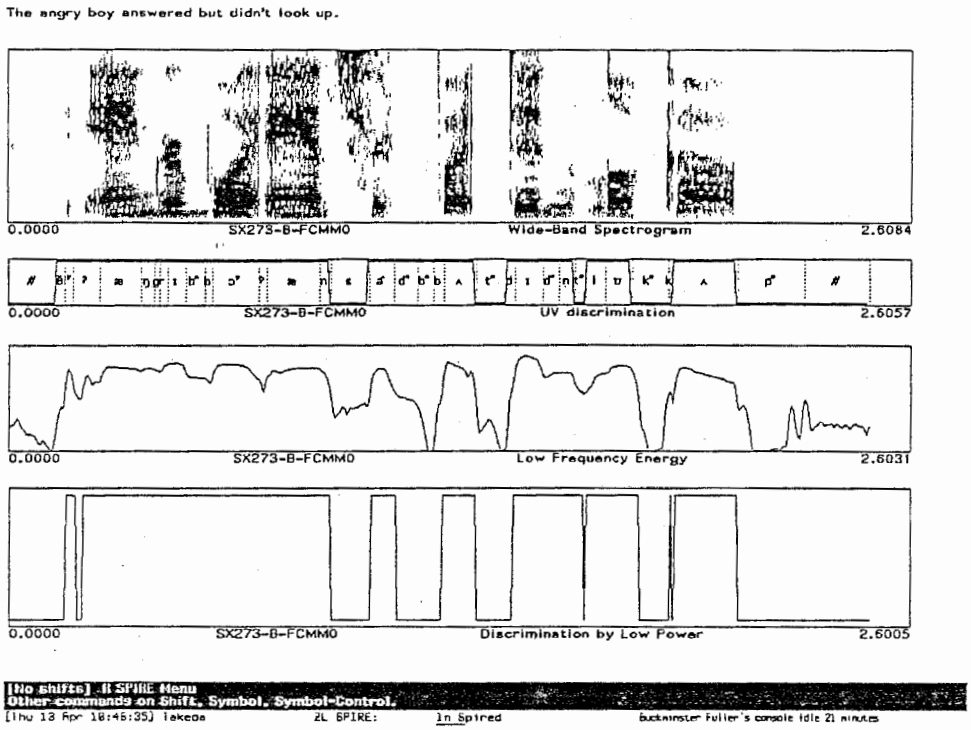

Figure 7: An example of the "voiced" feature detection using the "Low Frequency Energy" measurement (in the bottom window)

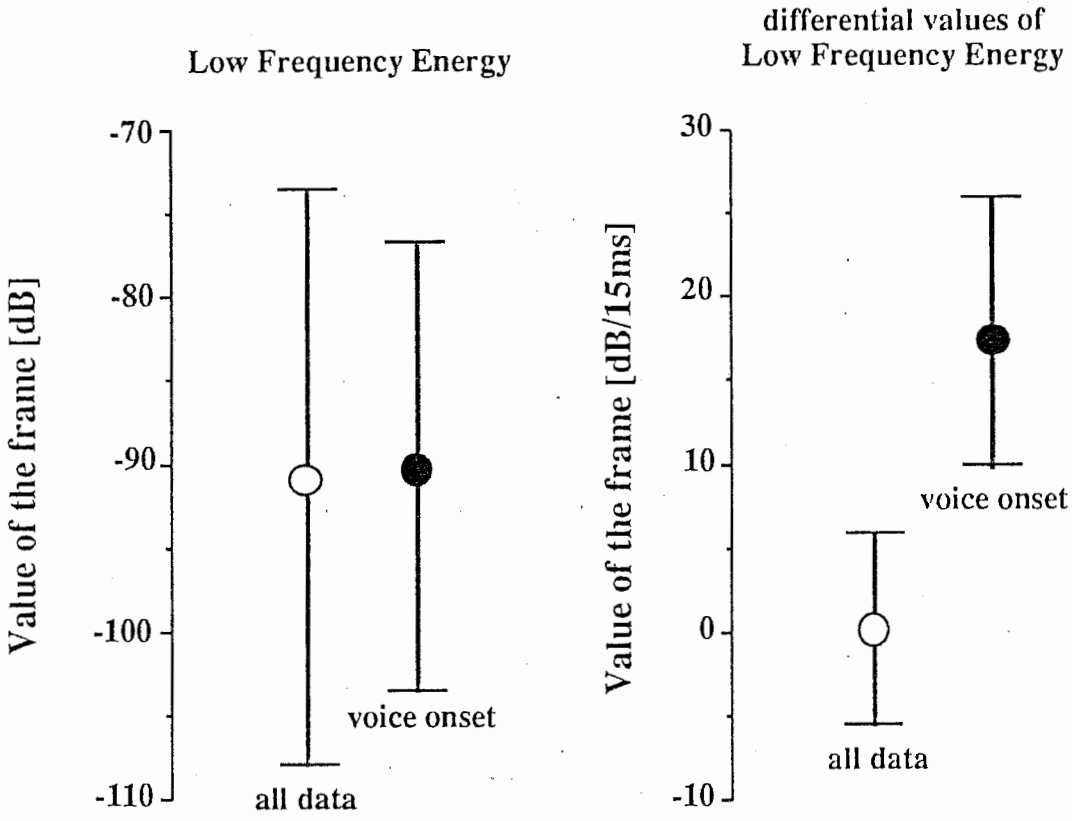

Figure 8: Distribution of the static and dynamic values

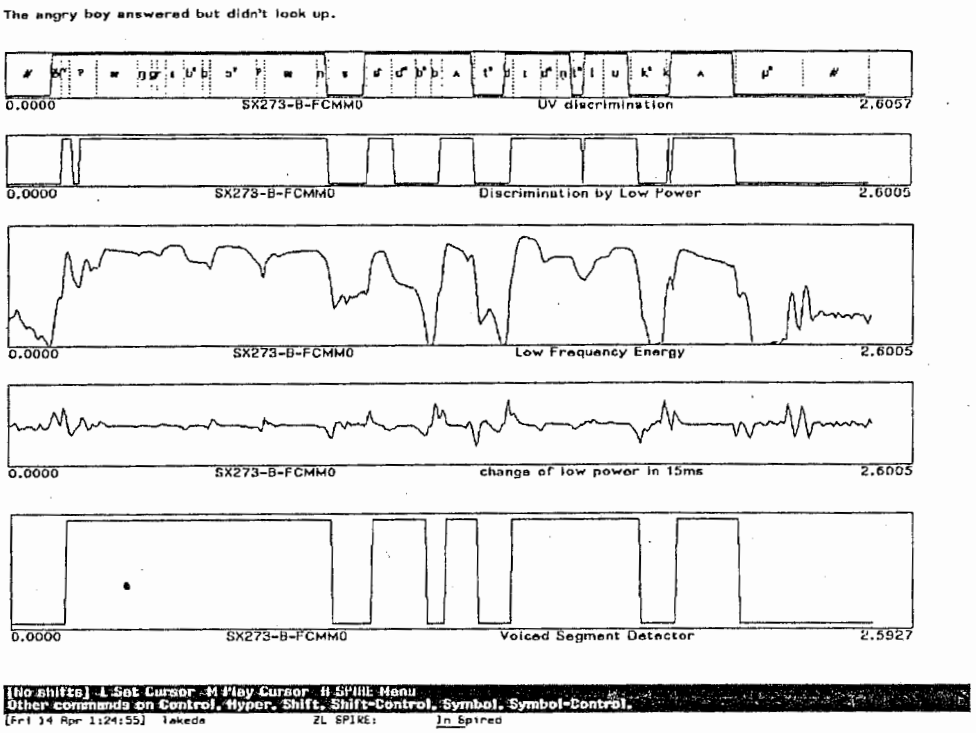

Figure 9: An example of the "voiced" segment detection by combining static and dynamic measurements (in the bottom window)