

## Phoneme Recognition Using Visual Features on Speech Spectrograms

Shigeru KATAGIRI\* and Manami YOKOTA\*

### ABSTRACT

In order to apply speech spectrogram reading heuristics to an automatic speech recognition system, a more accurate expression of the heuristics must be developed. In particular, the transformation between acoustic feature measurements and phoneme candidates must be developed in a quantitative manner.

In this paper, a visual acoustic-feature label and a phoneme identification approach using this label is proposed. The visual acoustic-feature label, which is a polygon on a speech spectrogram, represents some aspects of an acoustic feature by its own geometric characteristics. Preliminary experimental results show that phoneme identification using the visual acoustic-feature label is feasible for realizing the quantitative transformation rules between the acoustic feature measurements and phoneme candidates.

### 1. INTRODUCTION

Recently, speech spectrogram reading techniques have revealed that a speech spectrogram is rich with acoustic features which could be valuable in an automatic speech recognition system, and that human experts can read a speech spectrogram by using the heuristics about the acoustic features[1,2]. However, in order to apply these heuristics to an automatic speech recognition system, it is necessary to develop a more accurate acoustic feature detection algorithm, and more effective transformation rules between the acoustic feature measurements and the phoneme candidates.

This paper focuses on the transformation between the acoustic feature measurements and phonemes, especially consonants. This research is that it is strongly expected to develop accurate consonant identification rules for automatic speech recognition system construction. In order to investigate this problem, a visual acoustic-feature label which is accurately marked on a spectrographic feature is proposed. The final goal is to design the transformation rules in a quantitative manner, by analyzing the visual acoustic-feature labels in a large data base. Preliminary validation results of consonant identification based on the visual acoustic-feature labeling, using a limited data base are reported.

### 2. FEATURE DESCRIPTION

#### a. FEATURE DEFINITIONS

First, on the speech spectrogram, the energy-concentrated area and a closure, which is expected to reflect the articulation condition, is called the acoustic feature. The acoustic feature is based on speech production process knowledge and speech spectrogram reading heuristics. There are six kinds of acoustic features, i.e. formants, nasal-murmurs, buzz-bars, friction-patches, bursts and closures. Also, spectrographic values of the acoustic feature, e.g. formant frequency, buzz-bar duration and friction-patch energy, are referred to as acoustic feature measurements. Next, a phoneme segment is defined as follows. A consonantal phoneme segment basically means the portion where formants do not exist. In other words, it consists of only the nasal-murmur, the buzz-bar, the friction patch, the burst and the closure. Formants are included in a vowel phoneme segment. Finally, a phonetic feature described by combining several acoustic features is represented. If necessary, acoustic features of adjacent phonemes will be used to represent the phonetic feature. For example, /p/ is basically represented by the closure, the burst and the formants of adjacent vowels.

#### b. VISUAL ACOUSTIC-FEATURE LABEL

The visual acoustic-feature label is a polygon marked on the above-mentioned acoustic feature by hand-labeling. Table 1 shows six kinds of visual acoustic-feature labels, i.e. "formant", "nasal-murmur", "buzz-bar", "friction-patch", "burst", and "closure". The label notation corresponds to the acoustic feature name. The portions where these labels should be marked in principle are shown in the right column of this table. Examples of visual acoustic-feature labeling are shown in Fig. 1. If the acoustic feature appears clearly in a segment other than phoneme segments shown in Table 1, the label is marked on it. As illustrated, the visual acoustic-feature label represents the acoustic feature measurements and shows their shape, location, area, and direction.

#### c. CONSONANT FEATURE FRAME

In reading a speech spectrogram, it is well known that the place of consonant articulation is identified according to the transition direction of the lower formants in adjacent vowel segments[3,4]. Also, some studies have shown that formant transition due to consonant articulation is superimposed on the adjacent vowel formant transition[5]. As these results suggest, the identification of a consonant, sandwiched between vowels, needs to describe the acoustic feature dynamics across the Vowel-Consonant-Vowel (VCV) segment. Therefore, a VCV segment was chosen as a basic unit to represent phonetic features. A hypothetical segment, i.e. a silent segment, is applied to a missing vowel both at the beginning of a word and the end of a word.

A frame structure, called "consonant feature frame", is used to represent the consonant feature. As Fig. 2 shows, the consonant feature frame consists of a "preceding vowel" slot, a "following vowel" slot, and some acoustic feature slots. The frame name specifies a consonant. The "preceding vowel" slot and "following vowel"

\*ATR Auditory and Visual Perception Research Laboratories, TWIN 21 Bldg. MID Tower, 2-1-61 Shiromi Higashi-ku, Osaka 540 Japan.

slot specify adjacent vowels. The acoustic feature slot consists of a slot name, which represents an acoustic feature, and acoustic feature measurements which are derived from the visual acoustic-feature label. Examples of the acoustic feature slot values are illustrated in Fig. 1.

#### **d. CONSONANT IDENTIFICATION**

The consonant identification process can be divided into two steps. In the first step, VCV segments are roughly classified according to acoustic feature slot association. For example, a speech segment with only "closure" slot and "burst" slot could be classified into a voiceless stop, and another speech segments with one large "friction-patch" slot could be classified into a voiceless fricative. In the second step, these rough-classified segments would be quantitatively identified according to a discriminant function which is derived from a large data base.

Also, by applying the discriminant analysis to the acoustic feature measurements, the role of each visual acoustic-feature label in identifying consonants will be evaluated.

### **3. EXPERIMENT and DISCUSSION**

As a preliminary validation of consonant identification using the acoustic feature labels, consonant classification was performed using a limited speech data base.

#### **a. DATA BASES**

The visual acoustic-feature labeling was performed on Japanese phonetically balanced word utterances. These word sets were mainly composed of any sequence with two phonemes that can appear in Japanese. They were uttered by five female speakers and were sampled at 16 kHz. The visual acoustic-feature labels were marked on a wide-band computerized spectrogram.

#### **b. CONSONANT FEATURE FRAME STRUCTURE**

The same consonantal segment often bears different acoustic features. Also, structural differences can exist, i.e. a different association of some acoustic features, rather than the numerical difference of the acoustic feature measurements. The phonetic feature frame can represent such a structural difference in association with some visual acoustic-feature labels.

Figure 3 shows classification results according to the structure of the consonant feature frames. Bar graphs in this figure represent the ratios at which the consonant feature frames were categorized into classes according to their slot associations. For example, all /p/ have a burst slot only. While in /b/, the segments are grouped into three classes; the first class consists of a "burst" slot, the second, a "buzz-bar" slot, and the last class is comprised both a "burst" slot and a "buzz-bar" slot.

In the main, classification results showed that each consonant had the acoustic feature label association which was described by the articulation theory. As the articulation theory states, and as Fig. 3 shows, a voiceless stop has a burst, a voiced stop has a buzz-bar, an affricate has both a burst and friction-patches, etc. However, it is also evident that under actual speaking conditions, labels, other than those peculiar to the consonant, appear: for example, "burst & friction-patches" that were found in /s/ and /sh/. By using the ratios in Fig. 3, we can design a "confidence score" in a consonant identification rule.

#### **c. FORMANT TRANSITION**

Formant transition plays an important role in the speech spectrogram reading[3,4]. Also, a perceptual experiment demonstrated that the most important phonetic features concentrate around the portion of the utterance where the spectral variation is locally maximum[6]. Taking those results into account, the formant transition around the CV boundary bears rich information for identifying a consonant. Therefore, the transition direction of the two lower formants for every consonant was investigated.

Figure 4 shows the classification results of the second "formant" label across the boundary between a consonant and a following vowel. In this figure, the formant transition is represented as the transition value  $F2(20)-F2(0)$ , where the value,  $F2(0)$ , equals the frequency at the beginning of the second "formant" label and  $F2(20)$  represents the frequency at 20 msec after commencement. Consonants were grouped into three classes, i.e. "up", "flat", and "down" according to the above-mentioned transition value. An "up" class means that the transition value is more than 50 Hz, while the "down" class means that the transition value is less than -50 Hz. All other cases are classified into the "flat" class.

These results show that the second formant transition mainly reflects the place of articulation. Namely, most of the labial consonants in the data base have a rising F2 and most of the alveolar consonants have a falling F2. However, it was found that the typical formant transition based on the articulation theory does not always appear in utterances. With respect to the first formant, it was rather difficult to find a classification tendency peculiar to each consonant. However, in both cases, the main reason why there are so many exceptional phenomena is that co-articulation due to the phoneme environment influences the formant transition. Therefore, to apply formant transition characteristics in a phoneme identification, it is necessary to take the phoneme environment into account.

#### **d. DURATION**

It is well known that acoustic feature duration can yield rich information for identifying a consonant. For example, the duration of the burst and the friction-patch is valuable when classifying a voiced and voiceless consonants. Therefore, the durations of the visual acoustic-feature labels were investigated.

Figure 5 shows the average and the standard deviation of the durations for every visual acoustic-feature label. "Burst" label durations are shown for every stop. "Buzz-bar" label durations for voiced stops are totaled in the buzz-bar class. "Friction-patch" label durations are shown for fricatives and affricates, and as the "burst" labels of affricates are very short, they are included in the "friction-patch" label durations. The duration of /r/ indicates the "buzz-bar" label duration plus the "burst" label duration. These results illustrate that the acoustic feature tends to

have a duration peculiar to a consonant. As a general rule, the velar stop burst tends to be longer than the bursts of other stops, and the friction of the voiceless fricative tends to be longer than the voiced fricative. Using these results, we can quantitatively design consonant identification rules based on the acoustic feature duration.

Recent perceptual experiments demonstrated that each consonant has a duration range peculiarity which causes correct perception, and that the duration plays a more important role in stop perception than in other consonant perception[7]. Values in this figure support the perceptual results. The following ratios, i.e.  $(1-(\text{standard deviation})/(\text{average}))$  and  $(1+(\text{standard deviation})/(\text{average}))$  were calculated. Most of the acoustic feature durations are expected to be between these ratios. With respect to stops and affricates in the data base, these ratios were about 0.6 and about 1.4, respectively. These values are nearly equal to the duration expansion ratios which cause the correct perception of these consonants.

#### 4. SUMMARY

In order to describe the acoustic features of consonants in a quantitative manner, the visual acoustic-feature label and the consonant feature frame were introduced. Using a limited data base, some preliminary experiments were performed. The classification results using the phonetic feature frame quantitatively provided a ratio at which each consonant has a specific feature association. The results of the formant transition and duration described the quantitative aspects of the acoustic feature labels.

As results showed, consonant classification based on only one kind of acoustic feature label is not satisfactory. However, application of the acoustic feature label and the phonetic feature frame made it possible to describe the quantitative aspects of the phonetic feature. Applying discriminant analysis to these labels will provide more effective identification rules for a consonant.

#### REFERENCES

- [1] A.Cole(ed.); Perception and Production of Fluent Speech, Chapter 1, pp.3-50, L.E.A. Publishers, 1980.
- [2] V.W.Zue and L.F.Lamel; An expert spectrogram reader: A knowledge-based approach to speech recognition, ICASSP 86, Vol.4, pp. 1197-1200, April, 1986.
- [3] V.W.Zue, et al; Textbook of MIT Speech Spectrogram Reading, ATR, January, 1987.
- [4] R.Mizoguchi, K.Tsujino and O.Kakusho; A Continuous Speech Recognition System Based on Knowledge Engineering Techniques, ICASSP 86 Proc., Vol. 2, pp.1221-1224, April, 1986.
- [5] S.E.G.Ohman; Numerical Model of Coarticulation, J.Acoust.Soc.Am., Vol.41, No.2, pp.310-320, 1967.
- [6] S.Furui; On the role of spectral transition for speech perception, J.Acoust.Soc.Am., Vol.80, No.4, October, 1986.
- [7] S.Katagiri, Y.Tohkura and S.Furui; A role of duration in syllable perception, J.Acoust.Soc.Jpn., Vol.42, No.2, pp.97-105, February, 1986.

Table 1. Six visual acoustic-feature labels. Spectrographic portions where each label should be marked in principle are shown on the right.

label	spectrographic portion
formant	an energy concentrated area in a vowel segment.
nasal-murmur	an energy concentrated area in a nasal segment.
buzz-bar	• a voice-bar before an explosion in a voiced stop. • an energy concentrated area in a low band of a voiced fricative and a voiced affricate segment.
friction-patch	a noise-like energy concentrated area in a fricative and an affricate segment.
burst	a burst in a stop, an affricate, and a /r/ segment.
closure	a silent segment before an explosion in a stop and an affricate segment.

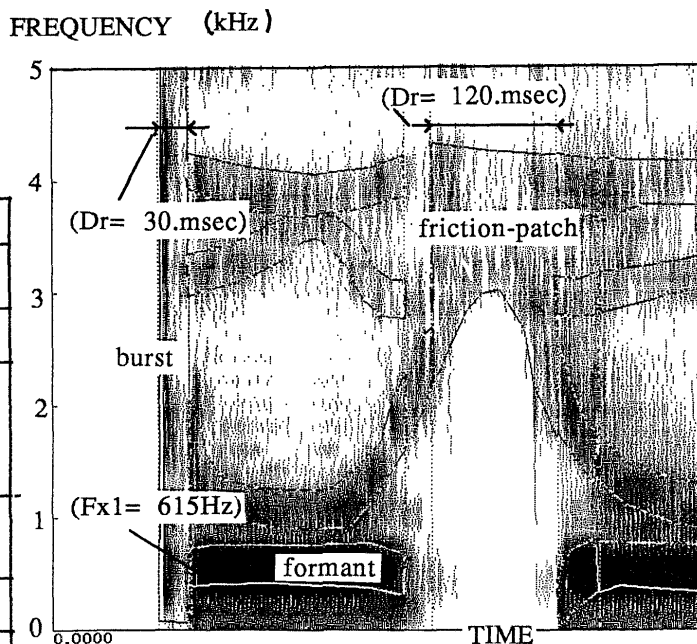
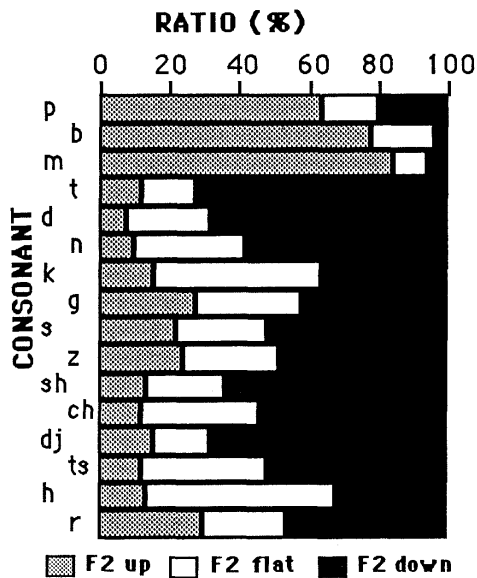


Figure 1. An example of a speech spectrogram and visual acoustic-feature labels. Three kinds of visual acoustic-feature labels and three phonetic feature frame slot values, which are parenthesized, are illustrated.

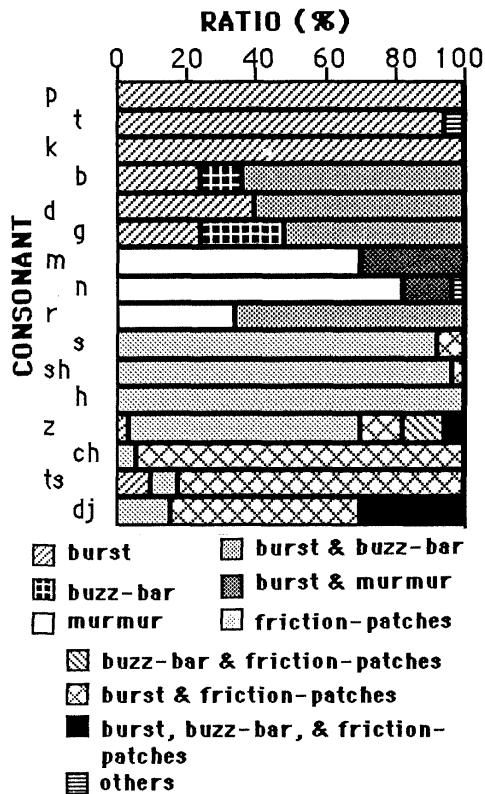
Consonant	
preceding vowel:	vowel
following vowel:	vowel
closure:	Dr
burst:	(Dr (E1 E2 E3 E4))
buzz-bar:	(Dr Fr Ea)
murmur:	(Dr Fr Ea)
friction-patch:	(Ct Ea (E1 E2 E3 E4))
preceding-vowel-formant:	((Fx1 Fy1)(Fx2 Fy2)(Fx3 Fy3))
following-vowel-formant:	((Fx1 Fy1)(Fx2 Fy2)(Fx3 Fy3))

Symbols in this figure stand for the following acoustic feature measurements.  
 Dr: duration  
 Fr: center frequency  
 Ea: energy  
 E1, E2, E3, E4: band-limited energy  
 Ct: cut-off frequency  
 Fxn: n-th formant frequency at the left measurement point  
 Fyn: n-th formant frequency at the right measurement point

Figure 2. The structure of the consonant feature frame. A slot in the frame consists of a slot name which represents an acoustic feature and its value which is derived from the visual acoustic-feature label. The slot name is shown left of the (: ) and the slot value is shown on the right.



F2(0) = formant frequency at the formant outset  
 F2(20) = formant frequency at 20 msec after commencement  
 "up" class:  $(F2(20) - F2(0)) > 50$  Hz  
 "down" class:  $(F2(20) - F2(0)) < -50$  Hz  
 "flat" class: others  
 Figure 4. Ratio of the second formant (F2) transition direction class. F2's were classified into three classes according to the transition value  $(F2(20) - F2(0))$ . The labels used here are based on utterances by 2 female speakers.



"friction-patches" means that there exist several friction-patches in a speech segment.

Figure 3. Consonant classification results according to the structure of the phonetic feature frame. Data is based on utterances by 2 female speakers.

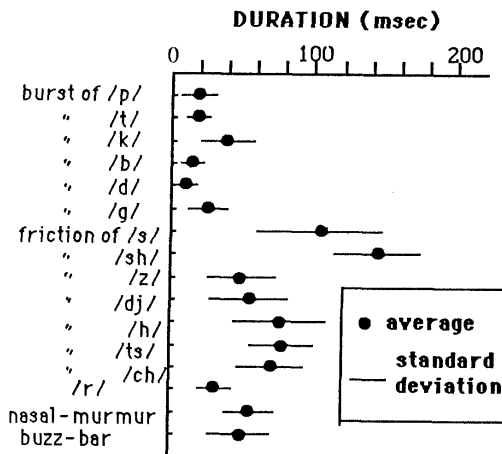


Figure 5. Durations of visual acoustic-feature labels. "Burst" label durations are shown for every stop. "Buzz-bar" label durations for voiced stops are totaled in the buzz-bar class. "Friction-patch" label durations are shown for fricatives and affricates, and as the "burst" labels of affricates are very short, they are included in the "friction-patch" label durations. The duration of /r/ indicates the "buzz-bar" label duration plus the "burst" label duration.