STUDIA PHONOLOGICA XXVII (1993)

An Experimental Study of Distinctive Features Using Speech Recognition Technology

Masatake DANTSUJI, Shuji DOSHITA and Shigeki SAGAYAMA

ABSTRACT

The present paper investigates the distinctive features using a recent developed technology for automatic speech recognition. This technology, "Phoneme Environment Clustering" (PEC), is formulated as an estimation of the mapping function from the "phoneme pattern space" which is a vector space corresponding to the phonetic(acoustic) aspect of the segment. The process of the successive splitting of sub-spaces forms a tree structure. Using PEC, we have examined approximately 2,000 segments from 216 phonemically balanced words uttered by a Japanese male informant. The results show that this tree diagram well corresponds to the phonetic "natural classes". The classification of segments into several groups reflects the differences between distinctive features. The results also show that the features [sonorant] and [consonantal] are separated from the others at the earliest stages of the process of constructing the tree structure. These acoustic hierarchies coincide with feature hierarchies based on articulatory properties in phonological feature geometry.

1. INTRODUCTION

In this paper we would like to describe an attempt to reconsider distinctive features and feature hierarchies by means of an advanced speech recognition technology. We have applied the speech recognition technology called "Phoneme Environment Clustering" for our experiment on distinctive features. Though the basic concept of distinctive features dates back some time, has quite early origins, one of the most important and epoch-making works on distinctive features is Jakobson, Fant and Halle's "Preliminaries to Speech Analysis" (1952). In the framework of this monograph, distinctive features are defined in terms of acoustic aspect as well as articulatory and auditory aspect. In the Jakobsonian framework, all distinctive features are binary and the basis of the distinctions should be taken as auditory or acoustic rather than articulatory.

Masatake DANTSUJI (壇辻正剛) : Part-time Lecturer, Department of Linguistics, Faculty of Letters, Kyoto University and Associate Professor, Faculty of Letters, Kansai University.
Shuji DOSHITA (堂下修司) : Professor, Department of Information Science, Faculty of Engineering, Kyoto University.
Shigeki SAGAYAMA (嵯峨山茂樹) : Supervisor, Speech and Acoustics Laboratory, NTT Human Interface Laboratories.

Part of this paper was presented at the XIIth International Congress of Phonetic Sciences at Aix-en-Provence, France.

The notion of distinctive features exerted great influence upon quite a lot of fields including generative phonology. In the early works of generative phonologists, they adopted the distinctive features of the Jakobsonian framework. Later, they revised their idea of distinctive features in many respects. The standard notion of generative phonology is well described in Chomsky and Halle's *The Sound Pattern of English* (hereinafter SPE) (1968). In the framework of SPE, distinctive features are mainly described from an articulatory point of view, and the same inclination has been maintained in current approaches.This, however, does not mean that acoustic and auditory aspects were assumed to have lesser importance, but rather that it was difficult to make an exact and precise description of the acoustic characteristics of distinctive features at that time. Jakobson et al. (1952) described the acoustic aspect of distinctive features on the basis of sound spectrographic analysis. We would like to examine an acoustic approach to the extraction of distinctive features using a speech recognition algorithm.

With respect to the hierarchy of distinctive features, several kinds of feature hierarchies have been proposed. For example, Fant (1973) argued for a feature hierarchy depending on economy of description in terms of the smallest number of features. Clements (1985) and Sagey (1986) proposed a feature hierarchy founded on phonological and phonetic aspects mainly based on the articulatory point of view. This tendency has continued into the domain of feature geometry in current phonology.

Recent experiments on speech recognition have revealed the possibility that there exists a different kind of hierarchy. For example, Dantsuji (1989b) describes a feature hierarchy depending on auditory distance between segments. We would like to introduce another kind of hierarchy based on acoustic distance in this paper. "Phoneme Environment Clustering" (hereinafter PEC), which was originally developed for automatic speech recognition, is here applied in an experiment aimed at establishing a feature hierarchy.

2. PHONEME ENVIRONMENT CLUSTERING (PEC)

As has been reported in many works, we can consider a number of possible factors concerned with a given segment, which may be affected by sound patterns of a given language, such as the preceding segment, the segment before the preceding segment, the center segment (the current segment itself), the succeeding segment, the segment after the succeeding segment, speakers, pitch frequency, power, speaking rate, stress position, phoneme position in the utterance, background noise, emotion and so forth. The combination of these factors makes an abstract space which is called an environment space E. Each allophone is assumed to be a point e in the space E. On the other hand, each allophone is also observed as an acoustic pattern which can be assumed to be a point v in a vector space (V) after some normalization of pattern durations.

If we have a set of phonetically labeled acoustic segments, each of them is a

point e in the environment space E as well as a point v in the pattern space V. Denoting the mapping function from the space E to V by $\phi: E \rightarrow V$, the acoustic pattern of each allophone $v=\phi(e)$ varies from sample to sample and has a certain spread in the space V. This spread is measured by some distortion measure, such as the averaged Euclidean distance from the centroid, and denoted by $d(v)$. The image in a subspace Ei of the phoneme environment space E acquired through the mapping function ϕ is also a subspace $Vi=\phi(Ei)$ in the vector space V. Its spread in V is denoted by $d(Vi)$.

The aim of phoneme environment clustering (PEC) is to find the optimal set of n subspaces $\{Ei\}_{i=0}^{n-1}$ to cover all variations of acoustic segments. It is d defined as the minimalization of the total distortion defined by:

$$D=\sum_{i=1}^{n} d\ (\phi(Ei))$$

where $\bigcup_{i=1}^{n} Ei=E, \quad Ei \cap Ej=\emptyset \quad (i \neq j)$.

That is, PEC aims to find an optimal division of the phoneme environment space to minimize the total sum of distortions of the images of the environment subspaces. The subspaces of E are also called phoneme environment clusters. This formulation means a sort of piecewise approximation of a mapping function such that, if an arbitrary phoneme environment is given, its pattern is predicted with a minimum error. Since it is not easy to obtain the real minimum, the solution to the above problem is approximated by successive splitting of the environment subspaces, which has the significant advantages that the clustering algorithm is simple, all produced subspaces are convex, the splitting process gives rise to a binary decision tree, and the binary tree is common regardless of the final number of clusters.

3. Experiments on PEC and Distinctive Features

As has been mentioned above, the process of the successive splitting of subspaces forms a tree structure which is interpreted as a similar relationship among phonemes and phoneme environment. The concept of PEC can be applied to the distinctive features, which are assumed to be components of phonemes, as well. For example, Fant (1973) stated that "the phonetic value of a distinctive feature can be regarded as a vector in a multidimensional signal space. The variability due to context shall be expressible by rules which define how the feature vector is changed when the conditioning elements are varied." Therefore, it is expected that distinctive features may be extracted from the experiment using the procedure of PEC to some extent.

The process of the successive splitting of subspaces forms a tree structure, as has been stated. We have examined how sets of phonemes are divided into allophones in the process of PEC.

Experiments were carried out under the following conditions.

1) Informants and texts: Approximately 2,000 segments out of 216 phonemically balanced words for one male adult and for one female adult, and approximately 6,300 segments out of 668 "bunsetsu" phrases from 100 sentences of an essay for one male adult.
2) Acoustic parameters: cepstrum, delta-cepstrum, log-power, delta-log-power.
3) Dimension: 34.
4) Regression window: 90 ms triangular.
5) Window length: 30 ms.
6) Window shift: 10 ms.
7) Sampling frequency: 12kHz.
8) Environment factors: 5 factors.
9) Distance measure: weighted Euclidean distance.

Fig. 1 shows an example of the tree structure which was formed through the process of successive splitting using PEC. It can be observed that allophones depending on phonetic environment are extracted at lower nodes. Phonemes as sets of allophones appropriately correspond to higher nodes which bind the lower nodes of the allophones.

Fig. 2 indicates the relationship between a phoneme, /k/, and its allophones. The phoneme /k/ is a voiceless velar stop and the precise place of articulation is influenced by the phonetic context. It is assumed that the influence of the following vowel is quite strong in the case of Japanese because of its syllable structure.

X-ray traces of the pronunciation of this phoneme indicate the precise difference in the point of articulation in accordance with the difference in the following vowel (Kokuritsu Kokugo Kenkyusyo, 1990) (see Fig. 3). When this phoneme is followed by the front vowels /i, e/, the point of articulation is somewhat advanced and is closer to the palatal position. On the other hand, when this phoneme is followed by the back vowels /u, o/, the point of articulation is somewhat retracted and is closer to the uvular position. When this phoneme is followed by the central vowel /a/, the point of articulation is neutral, viz. velar position.

From Fig. 2 we can observe the same tendency. The terminal node ③ is indicated as k (+k uo), and this means that the allophone followed by /u, o/ is separated at node ④. The symbol "+" means whichever phonemic environment is available in that position. In this case any phoneme can precede this allophone and this means that the influence of the preceding phoneme on this allophone is not very significant. The allophone followed by /a/ is separated in the next step and this is indicated by ②. The remainder is the allophone followed by the front vowel /i/, its corresponding semivowel (approximant) /j/ and another front vowel /e/. This is indicated by ①. The node indicated ④ itself is assumed to be the stage which extracts the phoneme /k/ as the bundle of these allophones.

Still higher nodes tie several phonemes into bundles. We would like to reconsider these groupings of phonemes by means of the notion "natural classes". It is well

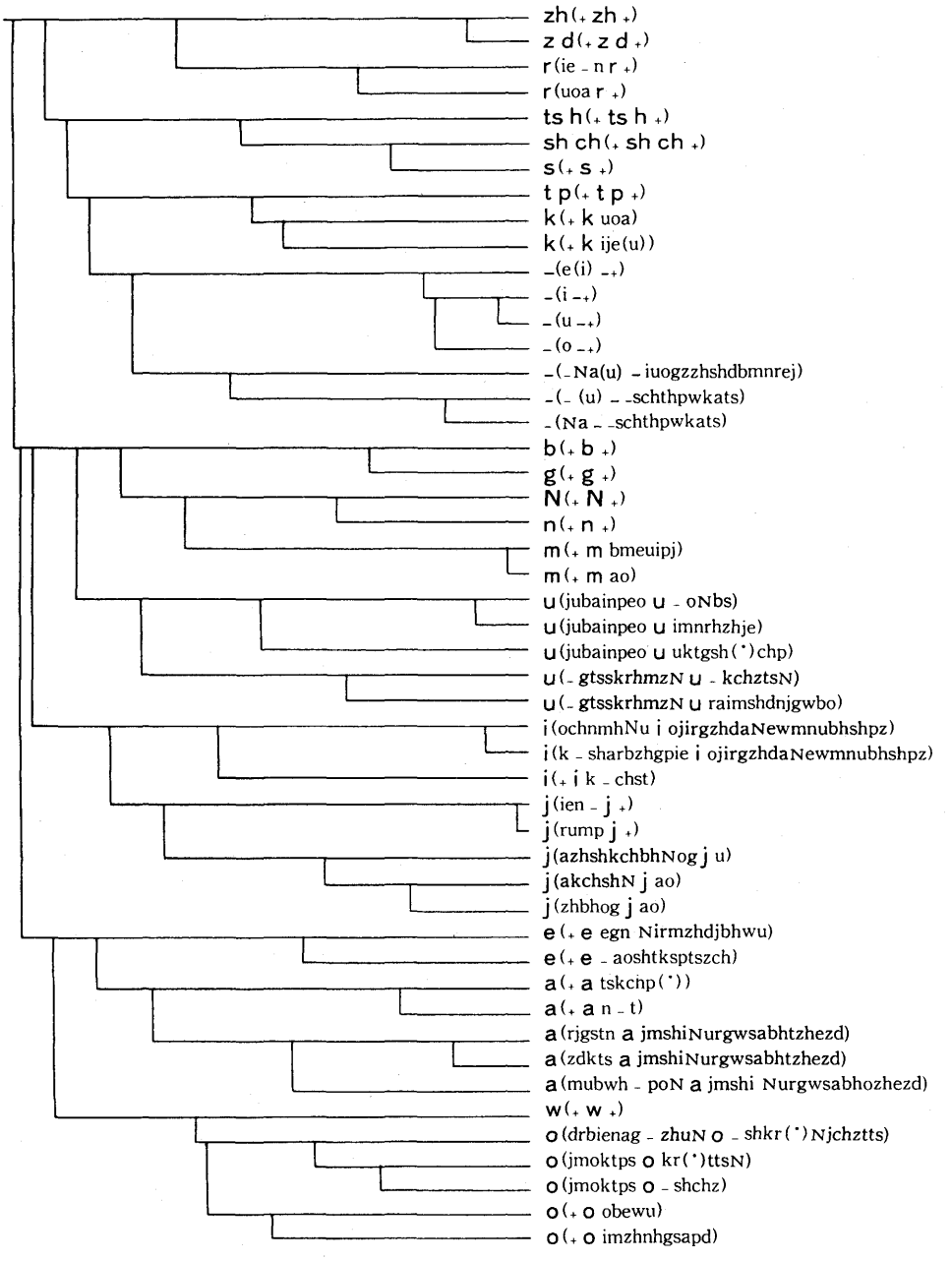


Fig. 1. An example of the tree structure resulting from PEC.

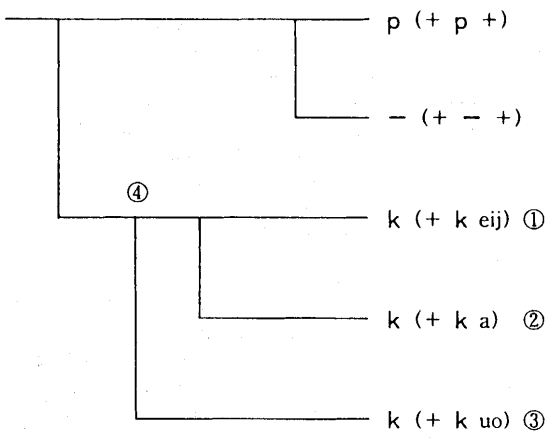
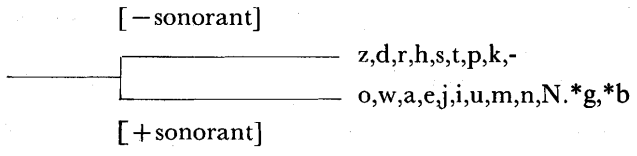


Fig. 2. The relationship between a phoneme /k/ and its allophones.

known that a set of segments called a "natural class" is formed in the field of phonology. Quite a lot of rules, not only synchronic but also diachronic, apply to sets of segments called natural classes. It is generally assumed that the extent of the naturalness of the class is in inverse proportion to the necessary number of the features which determine the class. However this barometer may not necessarily produce the correct result. Therefore, researchers have pointed out that phonetic validity is also required to establish a natural class. Phonetic distance between given segments in the phonetic space is one such notion and the result of PEC is expected to supply phonetic evidence of the acoustic distance for natural classes. Therefore, these bundles are expected to correspond to natural classes.

Fig. 1 shows that segments are divided into two groups in the first process, and this is interpreted as follows:



It may be observed from Fig. 1 that a set of segments which hold the feature [+sonorant] in common and a set of segments which hold the feature [–sonorant] in common are separated at the first step. The segment "h" is classified as one of the segments with [–sonorant] in this analysis. Sonorant sounds are generally defined as being produced with a vocal tract configuration sufficiently open that the air pressure inside and outside the mouth is approximately equal, and glides are usually classified as [+sonorant]. In regard to this point, however, there is a disagreement between several works on phonology. For example, Chomsky and Halle (1968) and

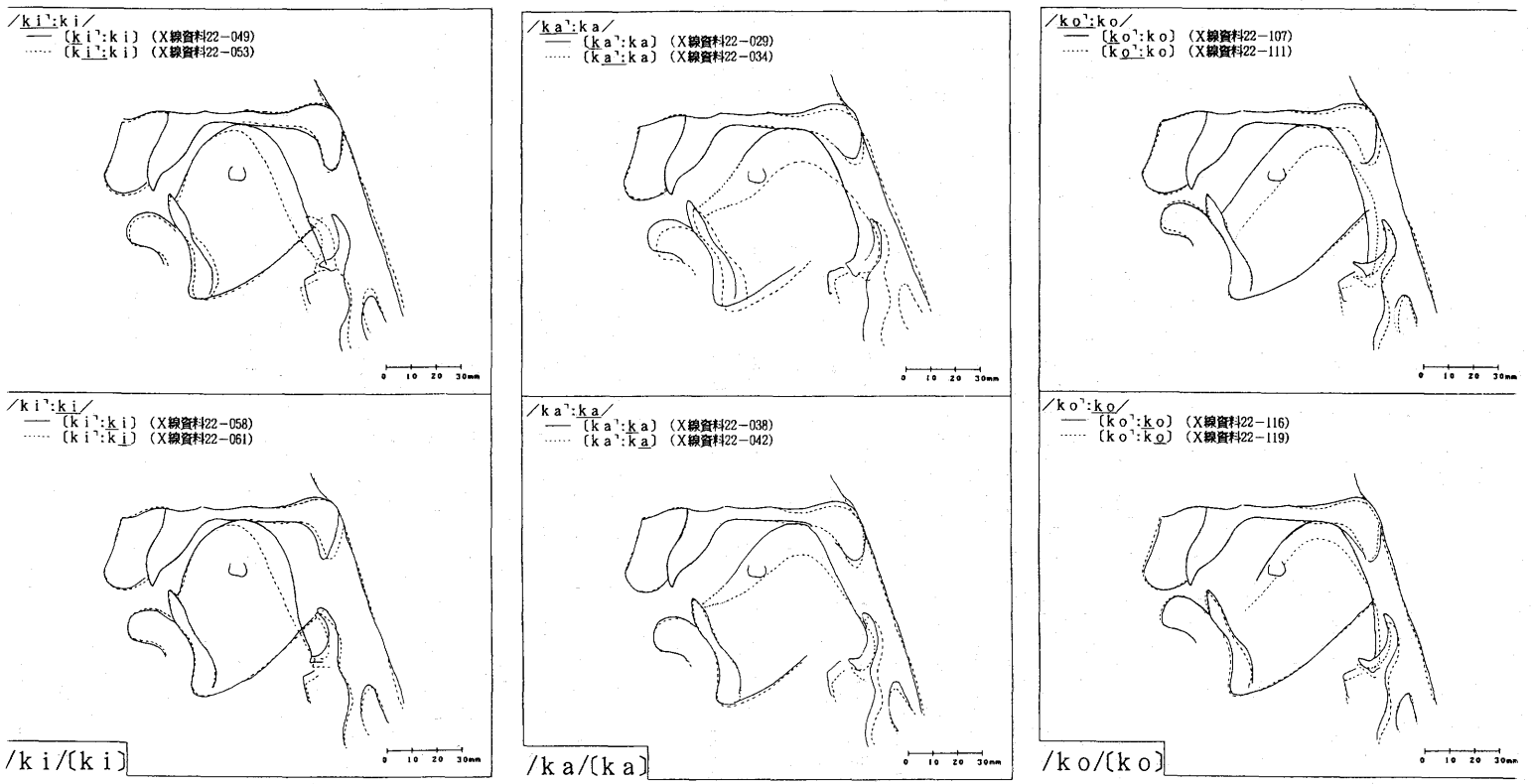


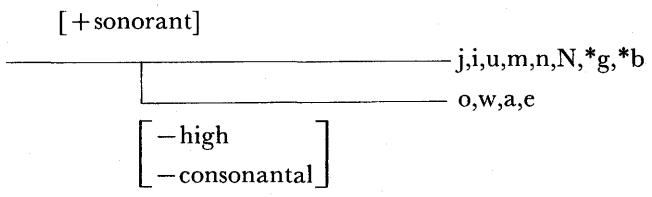
Fig. 3. Examples of X-ray traces of the phoneme /k/ from Kokuritsu Kokugo Kenkyusyo (1990).

Halle and Clements (1983) described "h" as [+sonorant]. On the other hand, Postal, who introduced the notion of the feature [sonorant], classified "h" as [−sonorant], and Schane (1973) also described it as "non-sonorant". Ladefoged (1971) did not define "h" as a sonorant sound, and, therefore, did not define it as a glide either. Fisher-Jørgensen (1975) considered this view an improvement. In the case of Japanese, the phoneme /h/ occurs as the allophones [Φ], [ç], [x], and [ɦ] in addition to [h], and this phoneme is not usually classified as a glide. Therefore, there is no problem in classifying this segment as [−sonorant].

The segment /r/ is represented often as an approximant (semi-vowel) in the case of English, and this would be classified as [+sonorant]. In the case of Japanese, however, this segment has numerous allophones and free variations. For example, /r/ is often represented as a kind of stop at word-initial positions, and as a flap or tap at word-medial positions. Carr (1993) assumes that taps are like "short" stops. However, at the same time, he assumes that the voiced alveolar tap [ɾ] is [−obs, −cont] whereas the voiced alveolar stop [d] is [+obs, −cont]. He mentions that this seems to undermine the notion that voiced taps are like 'short' stops since it denies that they are obstruents at all, but it does bring out the manner of articulation property ([−cont]) shared by stops and taps. He classifies taps as sonorants. However, if the taps are assumed to be "short stops", they should be defined as [−sonorant]. We, therefore, would like to classify Japanese /r/ as [−sonorant] on the basis of our acoustic analysis. This is supported by the fact that Japanese /r/s often appear as kinds of stop and tap or flap. Therefore, we classify this segment as [−sonorant] in this instance.

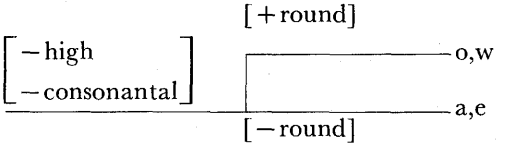
Attaching asterisk (*) to "g" and "d" implies a special sense here. These segments are in fact voiced stops and should be classified as [−sonorant]. However, at the stage of labeling by preconditions for the phoneme environment clustering, we did not include transition portions of formants in the vowels but in the voiced stops. Therefore, a part of the properties of vowels, which should be classified as [+sonorant], is assigned to these segments in this analysis. Furthermore, /g/ and /b/ seldom occur as the voiced stops [ɡ] and [b]. They rather occur as the voiced fricatives [ɣ] and [β] or the velar nasal [ŋ] called "bidakuon". These are also assumed to be factors that make these segments [+sonorant].

At the next step, the segments that have the features [−high, −consonantal] in common were separated from those with [+sonorant].

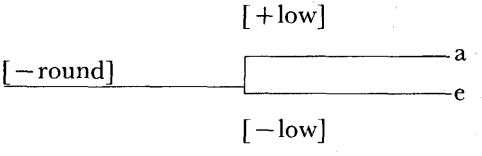


In this analysis /w/ is classified as [−high], although it is classified as [+high] in the case of English. In English or French, [w] is produced with a constriction between the upper and lower lips and the back of the tongue and soft palate as well, and is a so called voiced labial-velar approximant. On the other hand, in the case of Japanese, the degree of rise of the back of the tongue is lesser even at the word-initial position, and it is pointed out that the degree of rise is still lower at the word-medial position. The informant of this analysis reflects these properties of Japanese, and /w/ was classified as [−high].

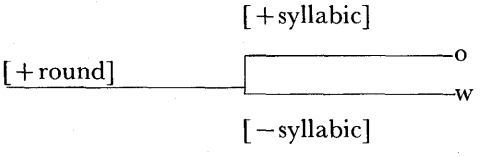
The group of segments which has [−high, −consonantal] is subdivided into a group which has the feature [+round], viz./o/ and /w/, and a group which has the feature [−round], viz. /a/ and /e/.



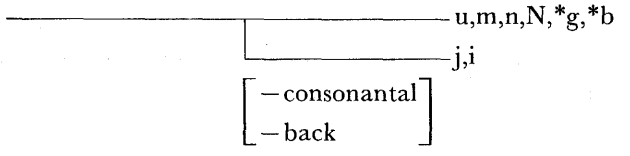
The segments that have the feature [−round] in common are further subdivided into the individual phonemes /a/ and /e/ by the feature [+/−low]. The low vowel /a/ and the non-low vowel /e/ are separated by this feature.



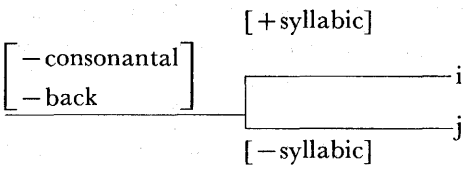
The segments that have the feature [+round] in common are further subdivided into the individual phonemes /o/ and /w/ by the feature [+/−syllabic]. The back vowel /o/ that can constitute a syllable peak and the glide /w/ that does not constitute a syllable peak are separated by this feature.



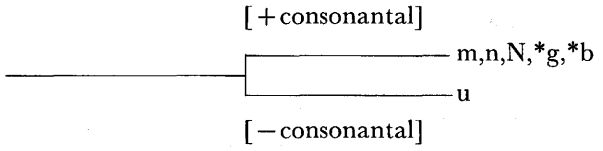
Other groups of segments are also classified and subdivided in a similar way.
The group of segments that has the features [−consonantal, −back] is separated from the remaining group of segments that has the feature [+sonorant].



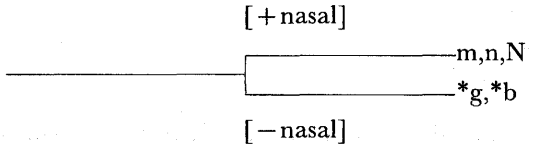
The segments that have the features [−consonantal, −back] are subdivided into the individual phonemes /i/ and /j/ by the feature [+/−syllabic]. The front vowel /i/, that can constitute a syllable peak, and the glide /j/ that does not constitute a syllable peak, are separated by this feature.



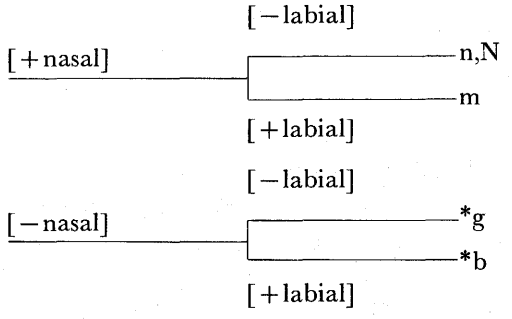
The high vowel /u/ is separated from the remaining segments by the feature [−consonantal].



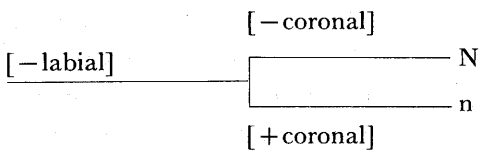
The group of segments that has the feature [+consonantal] is further subdivided into the groups /m,n,N/ and /*g,*b/ by the feature [+/−nasal]. The group of segments /m,n,N/ is characterized by the feature [+nasal].



The bilabial nasal /m/ is separated from the group of segments with [+nasal] by the feature [+labial]. The bilabial stop /*b/ and the velar stop /*g/ are also separated from the group of segments with [−nasal] by this feature [+labial]



The group with [−labial] is also subdivided into the individual phonemes /N/ and /n/ by the feature [+/−coronal].



Other groups of segments are also subdivided into individual phonemes in a similar way.

4. Feature Hierarchies

Recently, there has been a tendency to revise not only partial problems but also the total frameworks of feature systems in many ways.One of the main methods has been to set up a hierarchy structure or groupings for the feature arrangement. So far, several kinds of feature hierarchies or groupings of features have been proposed. For example, in a Jakobsonian framework, Fant (1973) discussed a feature hierarchy depending on economy of description in terms of the smallest number of features. From a study of automatic recognition study, Dantsuji (1989b) proposed a feature hierarchy making use of auditory distance.

In a generative phonology framework, for example, Clements (1985) discussed feature hierarchy geometrically organized from a phonological point of view, taking articulatory aspects into account. It is said that Sagey (1986) elaborated this feature hierarchy out of phonetic and physiological fact. Ladefoged proposed two types of features, auditory and physiological, the latter constituting a hierarchy. These phonetic and physiological facts mean that speech sounds are produced by the movement and action of a physiologically limited number of articulators, as was pointed out by Ladefoged and Halle (1988), McCarthy (1988), etc. The movable articulators are lips, tongue tip, tongue blade, tongue dorsum, tongue root, soft palate, larynx and so forth. Therefore, as the terminal features [high]. [back] and [low] relate to the movement of the dorsum of the tongue, they are dominated by the non-terminal node 'Dorsal'. The features [anterior] and [distributed] are dominated by the node 'Coronal', and [round] is dominated by the node 'Labial', in a similar way. Among non-terminal nodes, these movable articulators, viz. 'Soft Palate', 'Labial', 'Coronal', 'Dorsal', etc. are in the lowest positions. Among them, as 'Labial', 'Coronal' and 'Dorsal' are related to the place of articulation, these nodes are dominated by a higher node, '(Oral) Place'. Furthermore, the '(Oral) Place' node and 'Soft Palate' node are dominated by a still higher node, 'Supralaryngeal'. By contrast, major class features such as [sonorant] and [consonantal] are directly dominated by a root node which is the highest position of the hierarchy, or situated as special features that constitute the root node.

On the other hand, the present analysis by phoneme environment clustering establishes another type of feature hierarchy which reflects the acoustic distance between segments. Features such as [sonorant] and [consonantal] are extracted at considerably early steps in this experiment. For example, [sonorant] is extracted at

the first step of the clustering. These facts indicate that the acoustic distance between segment groups having the features [+sonorant] and [−sonorant] is considerable. Therefore, this confirms the view that the feature [sonorant] should be placed in a higher position in the feature hierarchy, as has been proposed in the current literature of non-linear phonology based on articulatory and physiological facts.

5. SUMMARY

We have examined distinctive features and feature hierarchies with the speech recognition technology called "Phoneme Environment Clustering" (PEC). The concept of PEC is explained and the tree structure resulting from our experiment using PEC is examined with reference to phonetic evidence, phonological issues and distinctive features. The results show that the tree structure as the outcome of acoustic analysis reflects the phonetic facts which are based on articulatory properties and corresponds to the feature theory quite well. The results also suggest the possibility of providing criteria for phonetic labeling of the speech database. We would like in future to collect more materials and factors for further experiments of this kind and to make inquires into the problem of the structure of feature geometry from an acoustic point of view.

ACKNOWLEDGMENT

We would like to thank Professor Tatsuo Nishida of National Center for Science Information Systems for supporting our research in various ways. Part of this study was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science and Culture, Japan.

REFERENCES

Carr, P.(1993): *Phonology*. Houndmills and London: The Macmillan Press.

Chomsky, N. and M. Halle (1968): *The Sound Pattern of English*. New York: Harper and Row.

Clements, G. N. (1985): "The geometry of phonological features", *Phonology Year Book* 2, 225–253.

Dantsuji, M. (1989a): "Onseigaku to on'inron", *Kouza Nihongo to Nihongokyoiku*, ed. O. Sakiyama, Vol. 11, 29–59, Tokyo: Meizi Syoin (in Japanse).

Dantsuji, M. (1989b): "A tentative approach to the acoustic feature model", *Revue de Phonétique Appliquée*, #91,92,93,147–159. Mons, Belgium.

Dantsuji, M. (1992): "Onsei on'in inritu," in *Nihongogaku o Manabu Hito no Tameni*, ed. F. Tamamura, Sekai Sisousya, Kyoto (in Japanese).

Dantsuji, M. and S. Kitazawa (1988): "A study on an acoustic feature model for speech recognition by machines", *Research Report* No. PASL 63–15–1, 1–20.

Dantsuji, M. and S. Sagayama (1991): "A study on distinctive features and feature hierarchies through "Phoneme Environment Clustering" (PEC)", *Proc. XIIth Int. Congress of Phonetic Sciences*, Vol.3, 190–193, Aix-en-Provence, France.

Doshita, S., T. Kawahara, Y. Mizutani, H. Kojima, M. Ishikawa and S. Kitazawa (1989): "Speaker-independent discrimination of Japanese consonant in isolated syllables using pair-wise discrimination method", *J. Acoust. Soc. Japan*, 45, 827–836.

Fant, G. (1973): *Speech Sounds and Features. Cambridge,* MA: MIT Press.
Halle, M. (1992): "Phonological features", *International Encyclopedia of Linguistics* Vol.3, ed. W. Bright, 207–212, Oxford: Oxford University Press.
Halle, M. and G. N. Clements (1983): *Problem Book in Phonology.* Cambridge, MA: MIT Press.
Harada, T. and H. Kawarada (1988): "Recognition of stop consonants by augmented feature space method", *IEICE Technical Report,* Vol. 88, No.91, 23–30.
Hawkins, P. (1984): *Introducing Phonology,* London: Hutchnson & Co.
Jakobson, R., C. G. M. Fant and M. Halle (1952): "Preliminaries to Speech Analysis: The Distinctive Features and their Correlates", *Technical Report* 13, *Acoustic Laboratory* MIT (Cambridge, MA: MIT Press. 1969)
Jakobson, R. and M. Halle (1956): "Phonology and phonetics", *Fundamentals of Language,* Part I, 4–51.
Kawahara, T., Y. Mizutani, S. Kitazawa and S. Doshita (1988): "Application of pair-wise discrimination model to Japanese consonant recognition", *Studia Phonologica* XXII, 83–93.
Kenstowicz, M. (1994): *Phonology in Generative Grammar,* Cambridge, MA: Blackwell Publishers.
Kitazawa, S. and M. Dantsuji (1989): "A study on speech recognition based on fine phonetic features", *Research Report* No.PASL 01–3–4.
Kokuritsu Kokugo Kenkyusho (1990): *Nihongo no Boin, Siin, Onsetu.* Tokyo, Syûei Syuppan.
Ladefoged, P. (1971): *Preliminaries to Linguistic Phonetics.* Chicago: University of Chicago Press.
Ladefoged, P. (1989): Representing Phonetic Structure. UCLA Working Papers in Phonetics, 73.
Ladefoged, P. and M. Halle (1988): "Some major features of the International Phonetic Alphabet", *Language,* Vol.64, 577–582 .
McCarthy, J.J.(1988): "Feature Geometry and Dependency: A Review", *Phonetica,* 45, 84–108.
Postal, P. M. (1968): *Aspects of Phonological Theory.* New York: Harper & Row.
Sagayama, S. (1987a): "Onso kankyô no kurasutaringu", *Nihon Onkyôgakkai Kôen Ronbunsyû,* 1–5–15 (in Japanese).
Sagayama, S.(1987b): "Onso kankyo kurasutaringu no genri to arugorizumu", *Densi Tûsin Gakkai Gizyutu Kenkyû Hôkoku,* SP87–86 (in Japanese).
Sagayama, S.(1989): "Phoneme environment clustering for speech recognition", *ICASSP–89,* 397–400.
Sagey, E. C. (1986): *The Representation of Features and Relations in Non-linear Phonology,* MIT Diss.
Schane, S. A. (1973): *Generative Phonology.* Englewood Cliffs, NJ: Prentice-Hall.