STUDIA PHONOLOGICA XVI (1982)

Discriminant Analysis of Burst Spectrum for Japanese Initial Voiceless Stops

Shigeyoshi KITAZAWA and Shuji DOSHITA

Introduction

The important predominance of stops requires to achieve a high performance level of stop recognition for automatic speech recognition (ASR). Stop consonants in Japanese occupy 19.2% of all phonemic occurrences including both vowels and consonants (Doshita, 1965). (Voiceless stop consonants occuring 13.7% amount more than twice frequencies of voiced stop consonants (5.5%)). In English, 31% of all consonant occurrences, and 18% of all phonemic occurrences are stop consonants (Mines, Hansen, and Shoup, 1978). Therefore, stop consonants are probably the most studied of all the consonant classes, both in the areas of speech acoustics and perception. Nevertheless, the implementation of stop consonant recognition components in current ASR system does not reflect the large number of acoustic stop features revealed through these studies.

In Japanese, typical ASR system identified voiceless stop consonants as one class, and did not discriminate the difference concerning place of articulation. This kind of approach can be justified in some extent, because the occurrences of individual consonant category are quite unbalanced, i.e., 0.2% for [p], 7.3% for [t], and 6.2% for [k]. Very low occurrences of bilabial stop reduce ambiguities incurred in ASR systems. However in the following study, we will not use these prior probability to evaluate features and decision algorithm.

In order to accomplish sufficiently high stop recognition rate, stop features must be studied instead of neglecting them. In this study, we intended to find features which are useful to discriminate stop consonants on the phonetic level. If we aimed at only final recognition score, we would apply some *ad hoc* technique, such as multi-templates, fixed following vowel context, or fixed speaker environment. We intended as much as general property independent of vowel context and speaker. For simplicity, initial consonant in CV utterances is our subject of study.

Recent studies on Japanese consonants are mainly conducted among several ASR systems which are designed based on the discrete time warp algorithm for restricted word recognition, moreover similar algorithm is also applied for CV syllable recognition. Basic assumption of these researches is that consonantal

Shigeyoshi KITAZAWA (北澤茂良) : Assistant, Department of Information Science, Kyoto University.
Shuji DOSHITA (堂下修司) : Professor, Department of Information Science, Kyoto University.

features mainly reside on the CV boundary region, that is well known formant locus or direction of formant transition rules which had been studied in speech synthesis research. These approaches are of course substantially vowel dependent, and it is naive to expect that the synthesis rules fully cover enourmous variety of feature vector obtained by simple analysis of natural speech. Therefore more advanced and practically experienced researchers adopt multiple templates approach in order to cover phenomenal variety of natural speech, although sufficient number of templates are required for a class of stop consonant. Those systems seem to have achieved fairely good result, however, features used are vowel dependent and some of them are speaker dependent, and do not reflect recent psychophysical results.

The possibility that spectra sampled at the release of stop consonants provide distinctive shapes for place of articulation has been noted in several investigations of natural speech. Halle, Hughes, and Radly (1957) have shown that spectral analysis of the burst in isolation give rise to three classes of patterns associated with the three places of articulation—labial, alveolar, and velar. Searle, Jakobson, and Rayment (1979) have also shown that acoustic events derived from spectra sampled in the initial few tens of miliseconds following the consonantal release can be used to separate stop consonants into categories according to place of articulation. The work of Fant (1960) in particular has attempted to characterize the distinctive patterns derived from short-time spectral analysis of stop-vowel utterances.

Blumstein and Stevens (1979) described at some length the theoretical considerations (based on Fant, 1960) which predict that initial stops produced at each of the three place of articulation in English should give rise to characteristically different release "onset" spectra. In few tens of milliseconds following the stop release the overall spectral shapes are, to the first approximation, unaffected by the vowel context, although onset frequencies of formant transitions will differ across vowels. These theoretical onset spectra were described by Blumstein and Stevens as diffuse and flat or falling for bilabial stop, diffuse rising for alveolar stop, and compact for velar stops. Thus, in the Blumstein and Stevens view, the onset acoustic information at the release of stop consonant, whether produced by noise burst, by aspiration noise, by the first few pulses of phonation, or by some combination of these, constitute an integrated acoustic cue to place of articulation that is invariant across vowel contexts.

As a consequence, Blumstein and Stevens (1979) developed three templates intended to capture the essential characteristics of the three types of onset spectra. In a study using natural adult speech in which onset spectra of stop consonant-vowel syllables were fitted with the templates, over 80% of the consonants were correctly categorized for place of articulation. In other work, Blumstein and Stevens (1980) have shown that listeners were able to reliably identify place of articulation for stop consonants in synthetic CV syllables on the basis of onset segments as short as 20–25 ms. These results were obtained for stimuli with or without initial bursts, and with or without moving transitions of second and higher formants. Blumstein and

Stevens have concluded that the short-time onset sepctrum constitutes the primary cue for initial place-of-articulation distinctions, and the formant transitions provide secondary, context-dependent cues.

In this study, using spectra of initial voiceless stop consonants in CV syllables from 28 adult male speakers we examined whether the invariant characteristics of the burst spectra which denote Blumstein's templates are also identified for Japanese, and these invariant features are effective for discriminating stops across vowel context and different speakers. Invariant stop consonant features are interesting not only for psychophysical researcher but also for ASR system engineer, since current complicated decision and learning processes of ASR systems will be replaced by a very simple decision process which uses some invariant features without adaptation.

I Acoustic Analysis

A. Methods

Twenty-eight male speakers (aged 22–35, university students in post graduate course, all native Japanese of Kansai districts) were asked to read each of syllables in natural speed and equal tempo. Each talker was allowed to practice reading a list until he felt comfortable.

The experimental utterances included all possible Japanese voiceless stops including /ti/ and /tu/ which are usually pronounced as [ci] and [tsu]. Table 1 lists the individual utterances both in phonetic symbol and Kana letters which were read by the subjects. The subjects were instructed to read チ and ツ as [ti] and [tu] not [ci] or [tsu], and to try few times to adjust the recording level and pronounciation speed. The list of utterances typed on a sheet in lexcial order of Japanese dictionary consisted of 35 voiceless and voiced stop consonant syllables including five vowels. The analysis results for voiced stops will not be reported here.

The utterances were produced in a silent room in front of a SONY ECM-220FA microphone and a Hitachi Lo-D D-77S tape recorder was utilized for the recordings.

Table 1. List of CV syllables which speakers were asked to utter, in both phonetic symbols and Kana letters.

a	i	u	e	o	ア	イ	ウ	エ	オ
pa	pi	pu	pe	po	パ	ピ	プ	ペ	ポ
ta	ti	tu	te	to	タ	ティ	トゥ	テ	ト
ka	ki	ku	ke	ko	カ	キ	ク	ケ	コ
ba	bi	bu	be	bo	バ	ビ	ブ	ベ	ボ
da	di	du	de	do	ダ	ディ	デゥ	デ	ド
ga	gi	gu	ge	go	ガ	ギ	グ	ゲ	ゴ

Speech waveforms were digitized directly from the monitor output of the tape recorder in order to avoid phase distortion. A JEIC 3118 low-pass filter with 70 dB/oct for anti-aliasing (the cut-off frequency was set to 8.9 kHz) and a DATEL DAS-250 16-channel 12-bit A/D converter of 4-microsecond sampling period were used for digitization. Through the A/D converter which was connected to a FACOM U-200 minicomputer at the common bus with direct access mode, speech data samples at every 54 microsecond (18.5 kHz sampling) were stored into a cartridge-disc of 2-megabyte in real-time for about 1 minutes continuously. Speakers were allowed to read a list in natural way so far as reading falls in the 1 minutes of real-time recording interval. Individual utterances of CV segment was interactively separated and filed on a magnetic tape with additional information concerning speaker name, age, date, and description of phonemes.

Individual utterances were input from a digital magnetic tape to a minicomputer and the waveforms were drawn on an X-Y plotter. From the initial 3500 samples of data (190 ms), waveforms at the release were sampled manually using a cursor. Smoothed spectrum was analyzed and displayed on the plotter (Fourier spectrum), while corresponding parameters were computed by linear prediction algorithm (LPC) and stored on a disk store. The time window for spectral analysis was around 20 ms of varying length guided by the visual observation. The smoothed spectrum was displayed as a plot with a resolution of 13 Hz. Formants corresponding to spectral peaks were computed from the solution of higher order polynomial equation by the Muller method.

Because of the laborious nature of this manual procedure, the measurements had to be restricted to the most crucial aspects of the stop consonants—the burst spectrum, the onset frequencies of F_2 and F_3 following the stop release and the steady frequencies of F_2 and F_3. Although the release spectrum are affected by the following vowel context (especially for velar stops), there are some invariant features included in the spectral patterns across vowels and speakers. The extracted raw data consisted of two sets of LPC alpha parameters for burst release and onset and formant frequencies for steady vowel. A set of fifteen utterances was produced by each of twenty-eight speakers.

B. Feature Analysis

We will consider only burst spectrum here after, since we believe some invariant features reside in this portion. The first problem is how to measure a burst spectrum, that is, time window, spectrum estimation method, and computation of spectrum. The second is feature extraction, that is, conversion of spectrum patterns into more effective set of parameters.

The initial stop consonant consists of burst, aspiration, vowel onset, and vowel. The burst part of consonant is the sudden release of articulators from a state of complete occlusion. The aspiration follows (the excitation of the vocal tract by

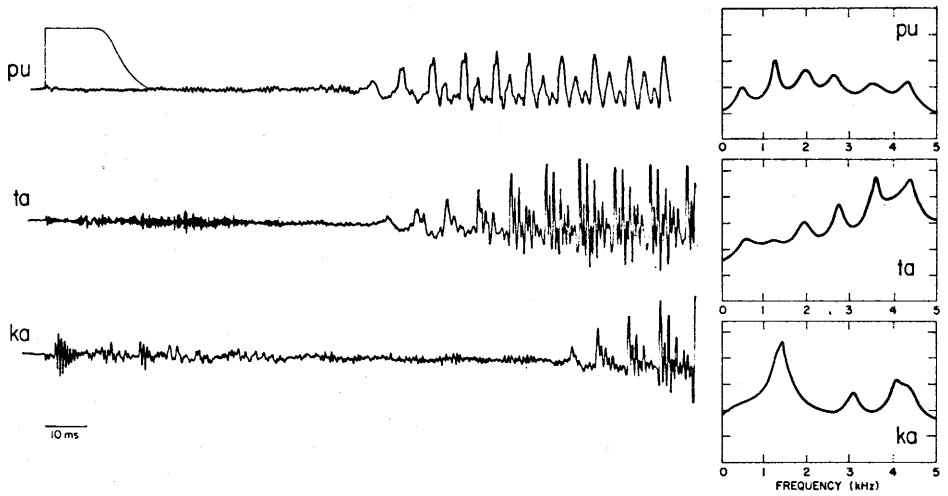


Fig. 1. Examples of waveforms and spectra sampled at the release of three voiceless stop consonants as indicated. Superimposed on the waveforms is the time window (of width 26 ms) that is used for sampling the spectrum by Blumstein and Stevens. (Reproduced and rearranged with permission, from Blumstein and Stevens, 1979).

glottaly generated turbulence).

The time window shape which Blumstein used, is shown in Fig. 1, puts greater emphasis on the earlier portions of the signal for a spectrum calculated at an onset. He chose a 26 ms time window because it seemed to produce spectral shapes that were optimally similar to the theoretically derived curves. Owing to the varying burst length and voice-onset time for these stops (also superimposed with speaker effects), the portion of the consonantal onset actually measured varies across the different consonants. Note that for [b], the 26 ms time window includes both burst and some portion of voicing onset, whereas for [g] essentially only the burst is measured.

We did not apply fixed window, though, Blumstein claimed fixed window is sufficient for a short time onset spectrum. Firstly, we intended to derive spectrum which is invariant across vowel as much as possible. Therefore our time window did not include vocalic portion. A 26 ms time window is too long for Japanese stops, since the aperiodic portion is not so long as Blumstein's observations. Japanese voiceless stops are weakly pronounced rather than English correspondents. Burst portion have to be carefully separated. In case of voiceless stops, the time window we applied includes the initial frication burst and possibly a portion of aspiration, but never extended into the onset of voicing. A varying length of time window (half-Hamming) positioned at the burst onset was used in deriving the spectra (see Fig. 2). Note that the aspiration portion is lightly weighted in order to emphasize burst spectra.

Examples of spectra for several naturally produced voiceless consonants in

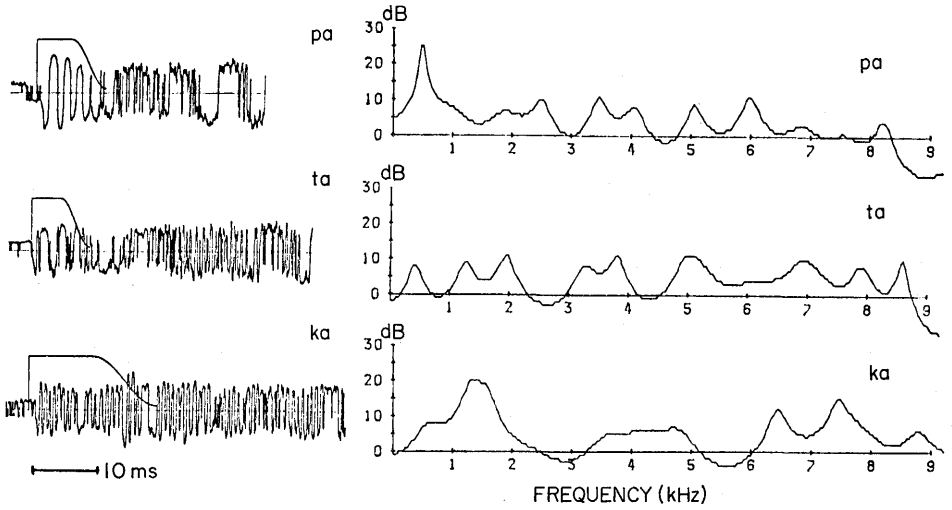


Fig. 2. Examples of waveforms and spectra sampled at the release of the three voiceless stop consonants as indicated. The amplitude of the waveforms is represented in sone scale, i.e., a power of voltage measured in linear scale with the exponent is set to 1/3 so as to emphasize the burst portion only for displaying purpose and not used in the following spectrum computation. Superimposed on the waveforms is the time window (half-Hamming) that is used for sampling the spectrum. Short-time spectra are determined for the first difference of the sampled waveform (sampled at 18.5 kHz) and are smoothed using a linear prediction algorithm, i.e., they represent all-pole spectra (26-th order) that provide a best fit to the calculated short-time spectra with pre-emphasis.

Japanese are shown in Fig. 2. These are linear prediction spectra obtained by pre-emphasizing the high frequencies and varying time window beginning at the consonantal release. The first difference of the waveform was calculated (in effect pre-emphasizing the high frequencies). A smoothed spectrum was calculated using a 26-pole linear prediction algorithm which is effective to estimate spectrum from very small number of time samples. Optimal prediction order was decided examining the final prediction errors, and resulted in a little higher order than vowels, since the burst portion should be represented by a pole-zero model instead of an all-pole autocorrelation model.

C. Spectral analysis

Although numerous methods have been proposed for representing the spectral characteristics of speech, such as band pass filter sampling and analysis-by-synthesis, there are a number of reasons why linear prediction techniques are becoming widely used (Markel and Gray, 1976). Because the model spectrum represents a smoothed version of the data spectrum with a very small number of parameters. The model used for representing the input data spectrum $|X[\exp(j\theta)]|^2$ is given by

$$\frac{\sigma^2}{|A(e^{j\theta})|^2} = \left|\frac{\sigma}{A(z)}\right|^2_{z=e^{j\theta}} \tag{1}$$

where σ^2 is a gain constant.

The spectral model is based upon the autocorrelation method. The first property of important is that on a log magnitude scale, A[exp(jθ)] or 1/A[exp(jθ)] for either the autocorrelation or covariance method has zero mean provided that A(z) has all of its zeroes within unit circle, i.e.,

$$\pm\int_{-\pi}^{\pi}\ln\left|A(e^{j\theta})\right|^2\frac{d\theta}{2\pi}=0 \tag{2}$$

In the autocorrelation method of linear prediction, the gain coefficient σ^2 is equal to the total squarred error α. The gain factor matches the energy or average value of the input spectrum $|X[\exp(j\theta)]|^2$ to the model spectrum $\sigma^2/|A[\exp(j\theta)]|^2$.

The model log spectra for the burst portion without gain factor are normalized spectra, will be used in the following discriminant analyses, since due to varying time window and recording level, the energy level did not account for discrimination.

We use the spectrum to denote model log spectrum from now on. A 26-th order spectral model with parameters $\{a_1, a_2, \ldots, a_{26}\}$ has been calculated, based upon input data sampled at 18.5 kHz. To have a frequency resolution of 30 Hz in spectrum, N' must satisfy $f_s(\text{kHz})/N' \approx .03$ or $N' \approx 333$. A spectrum is represented with 256-point fast fourier transformation. Choosing the closed power of two L=9 and N'=512 and a frequency resolution (distance between discrete samples) of 36.1 Hz. Using Fortran FFT subroutine to compute LM(1/A) at the discrete frequencies f_k=36.1 k, k=0,1,2,...,256, the Y array is filled with zeroes since the input sequences is real, and the X array is filled as follows:

$$\begin{array}{lcccccccccc} X = \{1, & a_1, & a_2, & \ldots, & a_{26}, & 0, & 0, & \ldots, & 0\} \\ & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \text{index} & 1 & 2 & 3 & 27 & 28 & 29 & & 512 \end{array} \tag{3}$$

After calling the FFT subroutine with L=9, LM(1/A) is computed from the Fortran variables X(J) (for real part) and Y(J) (for imaginary part) as

$$S(J)=10\log_{10}\left\{\frac{1}{\left|A\left(e^{\frac{j2\pi(J-1)}{N'}}\right)\right|^2}\right\}=-10\log_{10}[X^2(J)+Y^2(J)]$$
$$\text{for}\quad J=1, 2, \ldots, 257. \tag{4}$$

D. Feature selection

The selection of variables is key problem in pattern recognition and is termed as feature selection or feature extraction. Feature selection is considered a process of mapping the original measurements into more effective features. We have presented the burst spectrum by a 256-point log magnitude spectrum, however, this representation is sufficiently detailed, is too much redundant, because correlation between adjacent frequency components are high. Too many parameters cause computational complexity, singular matrix problem, and degraded computational precision. Theoretically one of efficient way of reducing dimensionality of feature space is to expand by principal components. Although this method reduces

redundancy as a whole, it does not evaluate effectiveness of feature set based on discrimination. We have tried this method and found that it is not so efficient provided small number of majour components are used for classification.

The critical bandwidth derived from the psychophysics of human auditory system is efficient representation of spectral features also for recognition system. We have compared several representations, i.e., original 256-point spectrum, equal bandwidth spectrum, critical bandwidth spectrum, and principal component representations, and found that the critical bandwidth was best with log magnitude power scale. In order to approximate critical bandwidth components up to 1 kHz, 10 segments were derived out of 30 components by averaging every 3 components, and from 1 kHz up to 9.25 kHz bandwidth of following components were merged so as to be proportional to log frequencies resulting 18 segments, hence amounted to 28 new components as a whole. This 28-dimensional feature vector representation for each burst spectrum was used as an input for the next decision process.

Although above mapping from spectrum to critical band spectrum is linear, in many application of pattern recognition, there are important features which are not linear function of original measurements, but are highly nonlinear functions for the given data. Such examples are second order statistics as covariance matrix or cross terms of two components. From the examination of variance-covariance matrices of each class of consonants, there were several components which showed prominently different values between two classes. Some of these components will be shown to be effective features for recognition.

From these features, our task is to select the features so as to maximize a criterion. This process is described in the following decision process as a variable selection step.

II. Decision Process

The final step in the phoneme recognition is to decide from the feature detector data what phoneme was uttered. Because we are obviously dealing with noisy data, an analysis based on statistical decision theory seems most appropriate.

First, we will state statistical classification methods have yielded successful classifiers. Second, the procedure by which a subset of the available features is selected for use in the classifier must be defined in the next section. If d is the number of features to be used in making the classification, then a pattern $(f_1,f_2,\ldots,f_d)$ is a point in E^d (d-dimensional Euclidian space) with f_i being the value of the i-th feature. A pattern correspond to the d feature measurements for a segment or frame of speech. A training set consists of N patterns x_n, $i \leqslant n \leqslant N$, which are used to "train" the classifier to identify the class correctly. A classifier's decision is said to be correct if it agrees with this listener's classification. In general, if patterns are drawn from R classes, then the objective of training is to devide the feature space E^d into R regions, with region r corresponding to class or category r, $1 \leqslant r \leqslant R$. The division of the feature

space into R region can be described by a set of R discriminant functions g_r, $1 \leqslant r \leqslant R$, where a pattern $\mathbf{x}$ is considered to be in class i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \text{for all} \quad j \neq i,\ 1 \leqslant i, j \leqslant R. \tag{5}$$

Different types of classifiers are obtained by making different assumptions about such factors as the statistical properties of the features and the form of the discriminant functions.

A minimum distance classifier is one discriminant method. Each class r for which g_r is minimum, $1 \leqslant r \leqslant R$. This corresponds to mean vector $\mathbf{m}_r$ and its $d \times d$ covariance matrix C_r. The discriminant functions to classify a pattern $\mathbf{x}$ are distance measure of the form

$$g_r = (\mathbf{x} - \mathbf{m}_r)^t\, C_r^{-1} (\mathbf{x} - \mathbf{m}_r) \tag{6}$$

where vectors are assumed to be column vectors, and superscript t denotes transpose. Pattern $\mathbf{x}$ is considered to be in the class r for which g_r is minimum, $1 \leqslant r \leqslant R$. This corresponds to a Bayes classifier, in which it has been assumed that the joint probability density function of the feature measurements for each class is a multivariate Gaussian distribution.

This classification method has the advantage of simplicity: given an adequate training set, all that is required to obtain the classifier is computation of the mean vector and covariance matrix for each class. Possible disadvantages of the method are the assumption that the probability density functions of the features are Gaussian, and the large size of the training set needed to produce an accurate estimate of the mean vector and covariance matrix for each class.

The Bayes liklihood ratio test has been shown to be optimal in the sense that it minimizes the expected cost or error. However, in order to construct the liklihood ratio, we must have the conditional probability density function for each class. In most applications, we must estimate these density functions using a finite number of sample observation vectors. Estimation procedures are available, but they may be very complex or require a large number of samples to give accurate results. Even if we can obtain the densities, the liklihood ratio test may be difficult to implement; time and storage requirements for the classification process may be excessive. Therefore, we are often led to consider a simpler procedure for designing a pattern classifier, leaving a finite set of parameters to be determined. The simplest and most common choice is the linear classifier.

When covariance matrices are equal for each class, that is $C_r = C$, $1 \leqslant r \leqslant R$, the discriminant functions are reduced to a set of linear functions of $\mathbf{x}$ as

$$g_r = -\mathbf{m}_r^t\, C^{-1}\, \mathbf{x} + \frac{1}{2} \mathbf{m}_r^t\, C^{-1}\, \mathbf{m}_r \tag{7}$$

The advantage of this method is robustness with respect to non-normality and unequal covariance matrices (Lachenbruch, 1975).

B. Stepwise Discriminant Analysis

We have chosen to use the discriminant analysis program contained in the

BMDP (Biomedical Computer Program-P) released by Hitachi Co. for M-series VOS3 system for this purpose (Dixon and Brown, 1977). The program BMDP7M permits stepwise discriminant analysis, i.e., direct categorization of all of the data, or establishment of category boundaries within a portion of the data (the "known" subset), then application of these boundaries to determine (i.e., predict) the phonemes in the reminder of the data (the "unknown" subset). BMDP7M performs a multiple group discriminant analysis. The variables used in computing the linear classification functions are chosen in a stepwise manner. At each step the variable that adds most to the separation of the group is entered. Both forward and backward selection of variables are possible; at each step the variable that adds most to the separation of the groups is entered into (or variable that adds the least is removed from) the discriminant function. By specifying contrasts we can state which group differences are of interest; these contrasts guide the selection of variables. For each case the group classifications are evaluated. Based on the posterior probabilities, a classification table is computed (prior probabilities can be specified for use in these computations). In addition a Jackknife-validation procedure can be requested to reduce the bias in the group classifications. The program computes canonical discriminant functions and plots the first two to give an optimal two-dimensional picture of the groups. Output includes means, standard deviations, F-statistics for distance between pairs of groups, and degrees of freedom for each variable at each step, Wilks' Λ (U-statistic) for multivariate analysis of variance, and Mahalanobis D^2 of each case from each group mean.

```
00010 PROBLEM TITLE='DISCRIMINATION EXCPT KITAZ & OKU (82 6 26)'./
00020 INPUT VARIAB=32.FORMAT='(4F2.0,28E12.5)'.UNIT=8./
00030 VARIAB ADD=11.USE=7,12,17,19,21,26,31,34,37,41,42,43.
00040        GROUPING=3./
00050 GROUP CODE(3)=1 TO 3.NAMES(3)=P,T,K.USE=1 TO 3./
00060 TRANSF X1=X( 5)*X( 5). X2=X(19)*X(19). X3=X1+X2. X(33)=SQRT(X3).
00070        X1=X(15)*X(15). X2=X(22)*X(22). X3=X1+X2. X(34)=SQRT(X3).
00080        X1=X(11)*X(11). X2=X(31)*X(31). X3=X1+X2. X(35)=SQRT(X3).
00090        X1=X(18)*X(18). X2=X(31)*X(31). X3=X1+X2. X(36)=SQRT(X3).
00100        X1=X(28)*X(28). X2=X(32)*X(32). X3=X1+X2. X(37)=SQRT(X3).
00110        X1=X( 7)*X( 7). X2=X(13)*X(13). X3=X1+X2. X(38)=SQRT(X3).
00120        X1=X( 8)*X( 8). X2=X(23)*X(23). X3=X1+X2. X(39)=SQRT(X3).
00130        X1=X(13)*X(13). X2=X(27)*X(27). X3=X1+X2. X(40)=SQRT(X3).
00140        X1=X(14)*X(14). X2=X(23)*X(23). X3=X1+X2. X(41)=SQRT(X3).
00150        X1=X(23)*X(23). X2=X(28)*X(28). X3=X1+X2. X(42)=SQRT(X3).
00160        X1=X(21)*X(21). X2=X(24)*X(24). X3=X1+X2. X(43)=SQRT(X3).
00170        X1=X(1) EQ 4. X2=X(1) EQ 5. X3=X1 OR X2.
00180        X4=X(3)+6. X(3) = X4 IF X3./
00210 PRINT NO STEP.NO POST.NO POINT/
00220 DISC ENTER=4.00.REMOVE=3.999./
00230 END/
00240 FINISH/
```

Fig. 3. Complete list of Control Language for BMDP7M which we used to classify [p], [t], and [k], and yielded 84.2% of correct recognition. The data and variables are described in Control Language instructions and X(3) is specified as the variable that classifies the cases into three groups—P, T, K. In TRANS paragraph new variables are generated as a set of sum square roots of pairs of input variables. The last two lines in the TRANS paragraph is for excluding data from 4th and 5th speakers which the recording condition was invalid due to hardware missetting. X(1) is speaker index, X(2) is practice number, and X(4) is vowel index.

C. Control language and listings

The method we have used is shown in Fig. 3 as a list of the control language. Three groups of stop consonants are [p], [t], and [k]. Variable X(3) is a grouping variable. Variables X(5)–X(32) are raw data of critical bandwidth spectrum. (Note that from here after the variable X(J) denotes the value S(J)/5.0 in section I.C.) Variable transformation is specified in the TRANSF paragraph, and X(33)–X(43) are generated. In the following forward and backward step, starting with no variable in the classification function, variables are entered and removed at a time according to the criterion specified. Variables are entered into the classification function if they have F-value larger than ENTER, removed from if F-value less than REMOVE.

III. RESULTS AND DISCUSSION

A. Mean vector and standard deviations

The mean vectors of each category obtained averaging zero mean vectors of each sample across 18 speakers (see Fig. 4(a)), shows the invariant characteristics of the burst spectra as Blumstein stated. Diffuse and flat for [p], diffuse and rising for [t], compact for [k]. Note that [t] mean vector has a peak at 5.3 KHz region, and [k] mean vector has peaks at 1.2 kHz region corresponding to second formant of back vowels and 4.7 kHz region corresponding to third and higher formants of front vowels. Probably because the plosion of Japanese bilabial stops is not as strong as that of English and has very short duration, the shape of the burst spectrum of bilabial stop is rather flat and not so low frequency peak prominent, and only a few samples have falling shape of spectrum as the Blumstein's template. Because alveolar stops are not followed by such a long aspiration as that in English, samples with rising shape of spectrum are not dominant.

The standard deviation of vectors are shown in Fig. 4(b). While both [p] and [t] classes have approximately uniform standard deviations, [k] class has dominant two peaks because two groups are combined together in one group. The multidimensional distribution functions for both [p] and [t] classes are approximately similar, but [k] class is a composit of two mode distributions corresponding to front and back vowels. Therefore classification functions need to reflect covariance at least for [k] class conditioned to treat [k] class as one group.

B. [p]–[t] discrimination

Discrimination between [p] and [t] class has been conducted and resultant classification functions and classification matrix and histogram of canonical variable were obtained (see Fig. 5). Variables entered into the classification functions after forward and backward stepping were X(7), X(17), X(21), X(23), and X(26), and Jackknifed classification score was 89.5%. Transformed variables were less effective

than original spectrum, therefore not entered into. The histogram of canonical value shows good discrimination.

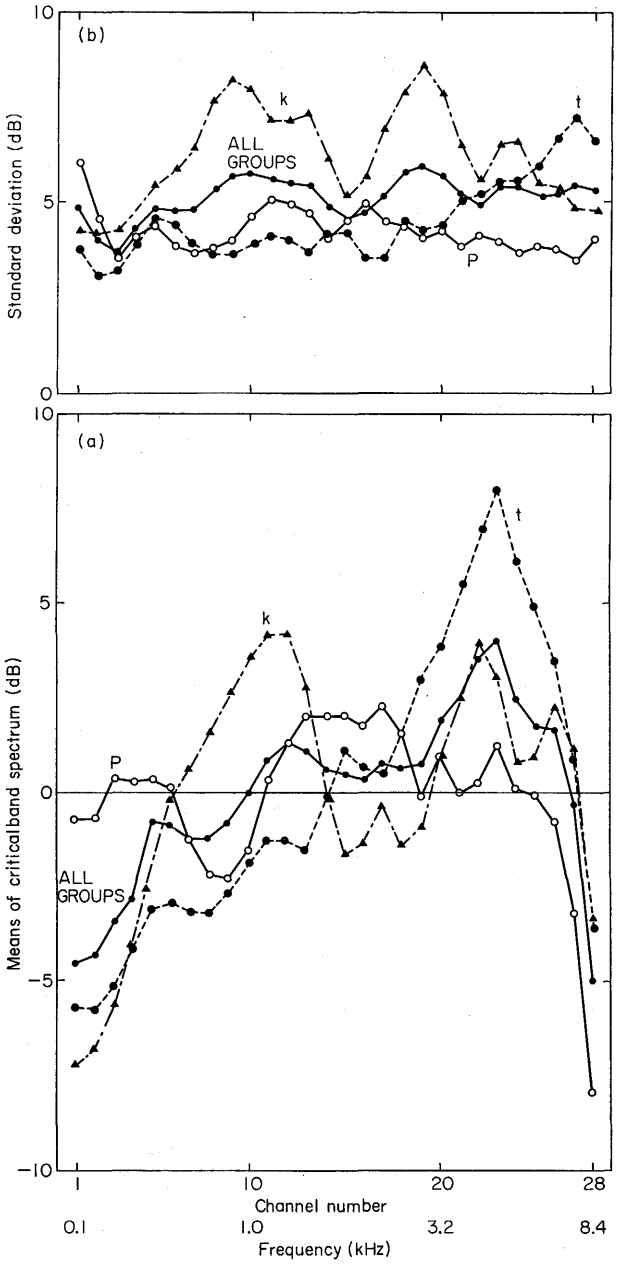


Fig. 4. Mean vectors (a) and standard deviations (b) for three category across samples from 18 speakers. The feature vector is critical band spectrum.

C. [t]–[k] discrimination

Discrimination between [t] and [k] class has been conducted and resultant classification matrix is shown in Fig. 6(a). Variables entered were X(5), X(12), X(17), X(19), X(21), X(27), and X(31), and Jackknifed classification was 81.6% correct which is a rather low score than [p]–[t] classification. Since the distribution of [k] class is different from the other classes (see Fig. 4(b)), it is necessary to consider cross terms of two variables. Variables entered were X(5), X(12), X(17), X(19), X(21), X(31), X(37), X(41), and X(43). Variable X(27) was removed and 3 transformed variables were entered. Jackknifed classification score rose to 87.1%.

CLASSIFICATION FUNCTIONS

	GROUP =	P	T
VARIABLE			
7	X(7)	0.87591	-1.96406
17	X(17)	0.57934	-0.43005
21	X(21)	0.85234	-0.28252
23	X(23)	-0.04259	0.38795
26	X(26)	0.42447	1.12993
CONSTANT		-1.08601	-2.72504

Fig. 5 (a).

CLASSIFICATION MATRIX

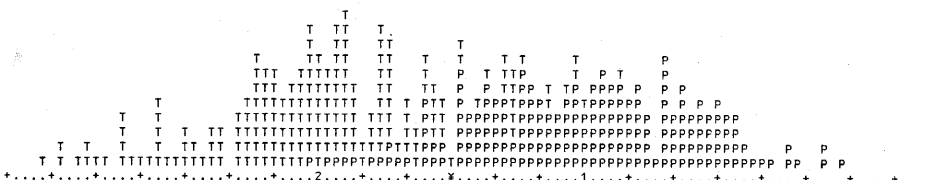
GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		P	T
P	89.5	170	20
T	89.5	20	170
TOTAL	89.5	190	190

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		P	T
P	89.5	170	20
T	89.5	20	170
TOTAL	89.5	190	190

Fig. 5 (b).

HISTOGRAM OF CANONICAL VARIABLE



ON THE AXIS ¥ INDICATES DIVIDING POINT AND NUMBERS THE GROUP MEANS

Fig. 5 (c).

Fig. 5. Results of [p]–[t] discriminant analysis by BMDP7M across 28 speakers utterances of three voiceless stop consonants. Classification functions (a), classification matrix (b), and histogram of canonical variable (c).

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		T	K
T	84.7	161	29
K	80.0	38	152
TOTAL	82.4	199	181

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		T	K
T	84.2	160	30
K	78.9	40	150
TOTAL	81.6	200	180

Fig. 6 (a).

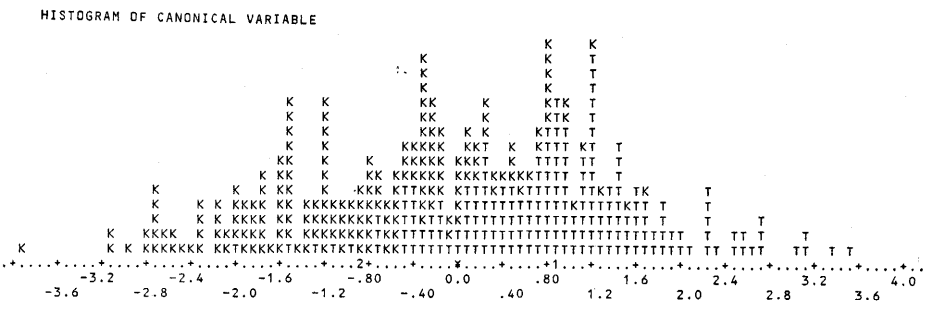


Fig. 6 (b).

Fig. 6. (a) Classification matrix of [t]–[k] discriminant analysis BMDP7M across 28 speakers utterances of three voiceless stop consonants. (b) Histogram of canonical variable.

Classification functions, classification matrix and histogram of canonical variable are in Fig. 7.

D. [p]–[t]–[k] discrimination

Variables were selected out of combination of variable set used in [p]–[t] discrimination and [t]–[k] discrimination, and resultant classification functions used X(7), X(12), X(17), X(19), X(21), X(26), and X(31). Jackknifed classification score was 77.0%, and dominant confusion occurred between [t] and [k] (see Fig. 8). Therefore we tried to improve discrimination by introducing cross terms. Finally entered variables into the classification function were X(7), X(12), X(17), X(19), X(21), X(26), X(31), X(34), X(37), X(41), X(42), and X(43). Note that newly entered cross terms were X(34) and X(37) which were effective for [p]–[t] discrimination, and X(41), X(42), and X(43) which were effective for [k] discrimination. The Jackknifed classification score was 83.3% for data of 28 speakers independent of vowel context. The improvement in correct recognition due to cross terms was statistically significant, furthermore the plots of canonical variables show more compact distribution of each class (compare Fig. 8(c) and Fig. 9(c)).

CLASSIFICATION FUNCTIONS

	GROUP =	T	K
VARIABLE			
5	X(5)	-1.78204	-2.36057
12	X(12)	0.26094	1.44018
17	X(17)	-0.21664	1.46730
19	X(19)	1.29858	0.48770
21	X(21)	-0.63856	0.15209
31	X(31)	0.01383	0.98129
37	X(37)	3.36407	2.62532
41	X(41)	1.35855	2.42677
43	X(43)	0.91230	2.25753
CONSTANT		-6.56025	-10.15124

Fig. 7 (a).

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		T	K
T	88.9	169	21
K	85.8	27	163
TOTAL	87.4	196	184

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -	
		T	K
T	88.9	169	21
K	85.3	28	162
TOTAL	87.1	197	183

Fig. 7 (b).

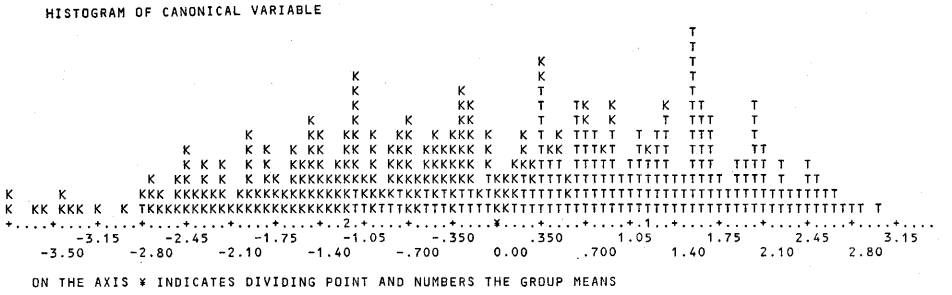


Fig. 7 (c).

Fig. 7. Results of [t]–[k] discriminant analysis by BMDP7M across 28 speakers utterances of three stop consonants using enhanced features including cross terms. Classification function (a), classification matrix (b), and histogram of canonical variable (c).

CLASSIFICATION FUNCTIONS

	GROUP =	P	T	K
VARIABLE				
7	X(7)	0.73966	-2.21307	-2.53393
12	X(12)	-0.96358	-0.31893	1.12619
17	X(17)	0.16826	-0.63041	0.60715
19	X(19)	-0.12377	0.20937	-0.70888
21	X(21)	0.20171	-0.92495	-0.02633
26	X(26)	0.06253	0.75377	0.78980
31	X(31)	-0.93626	-0.60977	0.30854
CONSTANT		-1.75494	-2.94150	-3.48165

Fig. 8 (a).

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -		
		P	T	K
P	86.3	164	21	5
T	75.8	21	144	25
K	72.1	2	51	137
TOTAL	78.1	187	216	167

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP -		
		P	T	K
P	86.3	164	21	5
T	74.7	22	142	26
K	70.0	2	55	133
TOTAL	77.0	188	218	164

Fig. 8 (b).

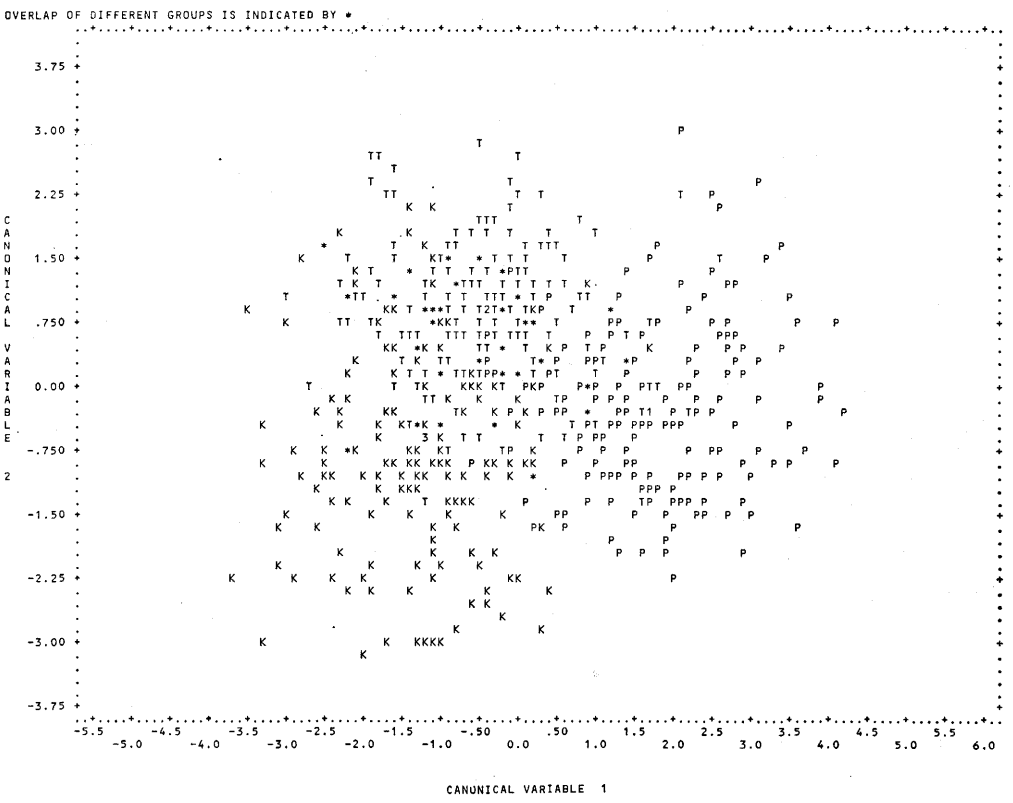


Fig. 8 (c).

Fig. 8. Results of [p]–[t]–[k] discriminant analysis by BMDP7M across 28 speakers utterances of three voiceless stop consonants. Classification functions (a), classification matrix (b), and plot of canonical variables (c).

CLASSIFICATION FUNCTIONS

GROUP = VARIABLE	P	T	K
7 X(7)	1.41562	-1.27225	-1.60248
12 X(12)	-0.64084	0.25155	1.59993
17 X(17)	0.64497	-0.03672	1.61363
19 X(19)	0.55729	0.95163	0.18325
21 X(21)	0.29110	-0.79347	-0.17860
26 X(26)	-0.13492	0.66397	0.55657
31 X(31)	-0.27537	-0.15901	0.76527
34 X(34)	1.24083	1.40784	2.16286
37 X(37)	3.31852	2.66767	1.94789
41 X(41)	1.14530	-0.01889	0.87371
42 X(42)	-0.58546	1.21333	1.30226
43 X(43)	1.45071	1.18489	2.57691
CONSTANT	-6.63453	-7.47735	-11.68442

Fig. 9 (a).

CLASSIFICATION MATRIX

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - P	T	K
P	88.4	168	18	4
T	81.1	20	154	16
K	83.2	5	27	158
TOTAL	84.2	193	199	178

JACKKNIFED CLASSIFICATION

GROUP	PERCENT CORRECT	NUMBER OF CASES CLASSIFIED INTO GROUP - P	T	K
P	88.4	168	18	4
T	79.5	21	151	18
K	82.1	6	28	156
TOTAL	83.3	195	197	178

Fig. 9 (b).

E. Speaker difference in discrimination

As we stated in section II, statistical method have to be carefully applied when the number of samples is not large. In order to validate previous results, a stronger test was applied. In computing classification function, one of speaker's utterances were excluded in order to reveal speaker's idiosyncrasies (Jackknifed classification was insufficient, since only a test sample was excluded and the other samples of the speaker under test were included in computation of the classification function). The 28-speaker data set was divided into a training set of "known" utterances and a test set of "unknown" utterances. The test set contained all utterances of one speaker and the training set contained remaining 27-speaker data set. The classification functions were obtained from the covariance matrix of 27-speakers, using this function the test set was discriminated. This test was repeated for each speaker's test data set in turn. The prediction run to discriminate [p]–[t]–[k] yielded 82.3% overall accuracy averaged for all speakers (see Table 2). Jackknifed classification degraded about 1% of accuracy than within class test, and "unknown" speaker test degraded also 1% than Jackknifed classification. Since these degradations were very small and the number of samples was not small (190 samples for each class),

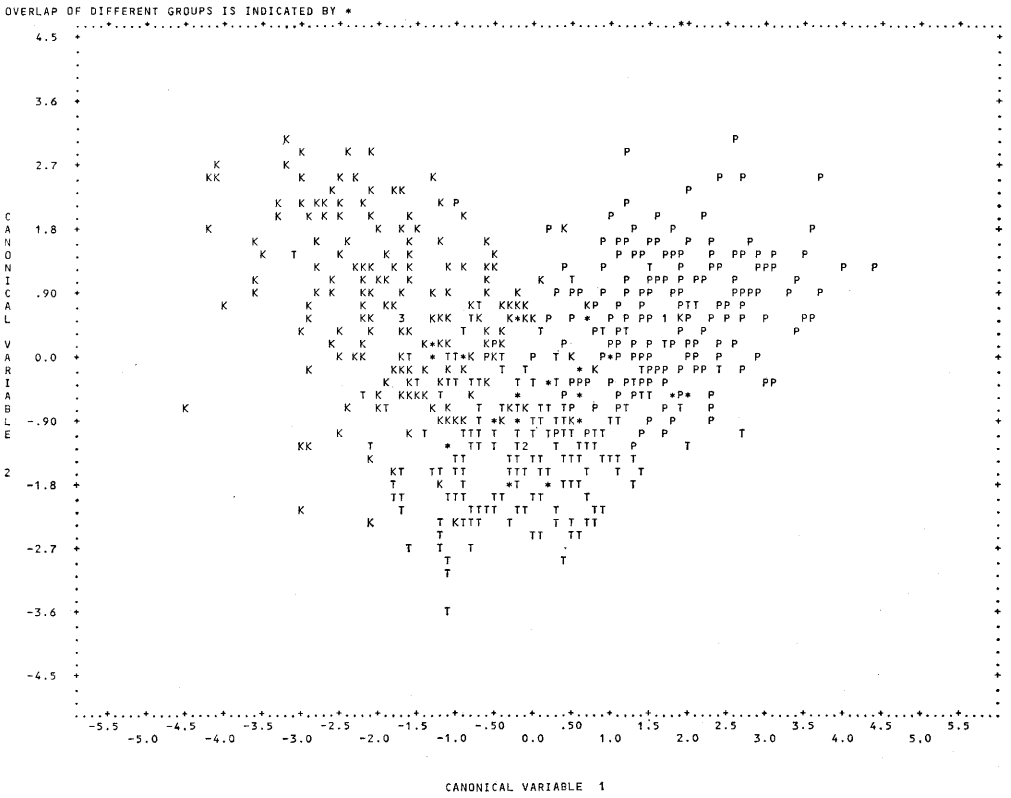


Fig. 9 (c).

Fig. 9. Results of [p]–[t]–[k] discriminant analysis by BMDP7M across 28 speakers utterances of three voiceless stop consonants using enhanced features including cross terms. Classification functions (a), classification matrix (b), and plot of canonical variables (c).

the estimated classification function will work about 80% accuracy for other unknown data set. Although we have to take account of the number of samples for each speaker, performance varied widely among speakers (e.g., 60% to 100%). The precise prediction of performance for individual unknown speaker was not possible now because of small number of samples.

F. Vowel dependent discrimination

The 28-speaker data set was divided into 5 groups according to the consonant following vowel. A group of data set contained 114 utterances of 28-speakers stops followed one of 5 vowels. Table 3 summarizes the results of discriminant analysis on [p], [t], and [k]. Table 3(a) is the confusion matrix resulting from a within-dataset classification of 114 utterances involving 28-different speakers. The program was able to classify 90.4% for vowel [a], 93.0% for [i] ,95.6% for [u], 86.8% for [e], and 92.1% for [o] of utterances correctly, using 12 of the features discussed in section D.

Table 2. Predictive classification of unknown 28 speakers utterances. Overall percent correct is 82.3%.

Actual phoneme	Speaker	Classified p	t	k	Speaker	Classified p	t	k	Speaker	Classified p	t	k	Speaker	Classified p	t	k
p	SG	10	4	1	MN	5	0	0	HR	5	0	0	OG	4	1	0
t		2	13	0		0	3	2		1	4	0		0	5	0
k		0	2	13		0	0	5		0	1	4		0	0	5
p	UC	15	0	0	SK	3	1	1	YS	5	0	0	HY	5	0	0
t		4	9	2		0	4	1		0	5	0		3	1	1
k		0	0	15		0	1	4		0	0	5		0	0	5
p	KB	13	2	0	NS	5	0	0	HS	5	0	0	YR	5	0	0
t		5	10	0		0	5	0		1	4	0		0	4	1
k		2	2	11		0	1	4		0	0	5		0	0	5
p	NR	12	3	0	SZ	4	0	1	YM	5	0	0	NY	5	0	0
t		1	12	2		0	5	0		0	4	1		0	3	2
k		3	0	12		0	1	4		0	1	4		0	0	5
p	NK	13	0	2	NG	2	3	0	KS	5	0	0	SD	5	0	0
t		0	15	0		1	4	0		0	4	1		0	5	0
k		0	3	12		0	0	5		0	1	4		0	2	3
p	MM	4	1	0	TM	3	2	0	KM	4	1	0	ST	5	0	0
t		0	5	0		1	4	0		0	5	0		0	5	0
k		0	2	3		0	2	3		0	2	3		0	3	2
p	MY	4	1	0	AR	4	1	0	MT	5	0	0	KT	5	0	0
t		3	2	0		0	3	2		0	4	1		0	3	2
k		2	0	3		1	2	2		0	2	3		0	1	4

Table 3. Classification of CV initial three stop consonants under fixed following vowel environments. (a) Within-dataset classification of known utterances. Overall percent correct is 91.6%. (b) Jackknifed classification. Overall percent correct is 89.5%.

(a) Classification															
Actual phoneme	Ca			Ci			Cu			Ce			Co		
C	p	t	k	p	t	k	p	t	k	p	t	k	p	t	k
p	33	5	0	37	0	1	36	0	2	33	5	0	34	4	0
t	5	33	0	0	35	3	0	38	0	6	30	2	4	34	0
k	0	1	37	0	4	34	3	0	35	1	1	36	1	0	37
Percent correct	90.4			93.0			95.6			86.8			92.1		
(b)															
p	32	6	0	36	0	2	36	0	2	33	5	0	32	6	0
t	5	33	0	0	35	3	1	37	0	8	28	2	4	34	0
k	0	3	35	0	5	33	4	0	34	1	2	35	1	0	37
Percent correct	87.7			91.2			93.9			84.2			90.4		

Of more significance are the results in Table 3(b). Here the classification functions were derived from all the data except the case being classified. This prediction run yielded 89.5% of accuracy. The errors are not uniformly distributed but instead are mostly in [e]. Of cource the number of samples for a group of vowel may not be sufficient to test a hypothesis of difference with statistically significant level.

IV. General Discussion

While the overall performance of the discriminant analysis averaged over the three places of articulation was 84.2% (82.3% for unknown speaker test), as compared to the Blumstein-Stevens (1979) result of 84.2% for voiceless stops, this comparison is very good coincidence. Although, Lahiri and Blumstein (1981) found the cross-language difference in French and Malayalam stops, the gross shape of onset spectrum we measured for Japanese stops differ from Blumstein's templates, the invariant property for place of articulation in stop consonants was also shown with more than 80% correct identification score across vowels and speakers. Since, for example, velar spectrum peak in the mid-frequency shifts depending the following vowel, velar stops are usually dealt in vowel dependent environments. In case of vowel dependent analysis, the averaged result over the five vowels was 89.5%, hence in some sense vowel dependent, but this dependency differs from vowel to vowel, [u] 93.9% was most dependent and [e] 84.2% was most independent.

The results which BMDP7M, a one of common program package for statistical analysis, computed and validated by Jackknife classification and by "unknown" speaker tests were found very stable. BMDP7M uses linear classification functions which is suitable and recommended in application fields, although the Bayesian decision is theoretically best. Further the disadvantage of linear classifier in classification power was compensated by inclusion of transformed variables.

The good performance achieved in this study is probably due to enhanced features in four points. First, spectrum extracted from wider frequency range of 9.25 kHz than usual 5 kHz range was effective for discrimination, since in the variable selection steps in BMDP7M some of higher frequency components were selected and contributed much for discrimination. Second, time varying window was adopted to select precisely the burst portion while sacrificing frequency resolution for compensation. However, we have no comparing data to Blumstein's fixed 25.6 ms time window. Third, the burst spectrum was averaged critical bandwidth wise, since we have found this representation is better performance in discriminating the place of articulation than other representation such as, original LPC log spectrum (256-point), uniform bandwidth spectrum, or principal components of spectrum. Fourth, cross terms that were designed as contrasts between within-class covariance matrices, were very effective in discriminating velar stops. There are of course other well known features of stops, i.e., formant locus or direction of transitions, however, none

of them results more than 80% performance when used under the vowel context and speaker independent environments.

Differences from other stop consonant studies have to be discussed. Our spectrum measurement is similar to Blumstein's method, but we used varying time window length because of above mentioned reason. Decision process is completely different. We used simple statistical procedure while Blumstein's procedure is descriptive and difficult to implement. Searle et al. (1979) used ciritcal band spectrum transformed into abstract features including formant transition accomplished 80% correct overall voiceless stops for unknown data set, 77% including detection and classification errors. The discriminant analysis program contained Statistical Package for Social Studies was used for decision process. Its high performance was due to combined spectral and transitive features and statistically optimized decision process.

In Japanese, direct comparison is possible, several researches were reported recently. Kobatake and Noso (1980) used majour two components of the principal component analysis on the burst spectrum represented by uniform 470 Hz bandwidth filter bank. Vowel dependent Bayesian classification of voiceless stop consonants resulted 78% correct recognition on unknown utterances. Degraded performance was probably due to information loss in feature reduction process. Tanaka (1981) using spectral peak locus in 50 ms after release of stop as a feature set, and representing typical patterns in stop consonant category by 14 potential functions, yielded 84% correct recognition of voiceless stops for 4 male speakers. These features are essentially vowel dependent. Ide, Honma, Makino, and Kido (1982) using 29 channel 1/6 octave filter output of five frames of every 10 ms, yielded 72% for burst spectrum, 90% for successive four frames. Direct comparison is difficult with our 76% for unknown without cross term result, since they included [p, t, k, c]. Mikami and Ohba (1981) using 3 pole LPC analysed burst spectrum to discriminate [p, t] and [k] for 100 speakers, i.e., contrasting diffuse and compactness, yielded overall 92.2% correct recognition of unknown dataset under the condition that the following vowel is known, however too much simplified feature limit final recognition.

From comparison of these results, our results was good, although we used only burst spectrum as features. However, it is true that features after burst portion or vowel onset to transition seem also possible to accomplish almost same recognition rate as shown by Searle and Tanaka.

The problem remained is how to discriminate remaining equivocal 20%. The invariant hypothesis is approximately true but has limitation of itself, as we have already shown, due to dependency on vowels and speakers. Among our database, some of burst spectrum distributed quite close to other center of category and their liklihood were as high as more than 90% of predictive probability. Recently Blumstein, Isaacs, and Mertus (1982) claimed that the gross shape of the onset spectrum may contribute to phonetic decision, it did not provide the primary

perceptual attribute to identify place of articulation. An opposite conclusion against his previous studies! According to his perceptual test of synthetic plosives, the majority of responses corresponds to the onset formant frequencies appropriate to the particular phonetic category.

Secondary cues such as directions of formant motions or frequencies of particular formants at consonantal release also provide significant information with regard to place of articulation (Blumstein and Stevens, 1980). This use of secondary cue is most clearly seen when the primary attributes of the onset spectrum are equivocal, so that the spectrum does not demonstrate strong unambiguous properties such as compactness or diffuseness, or is neutral with respect to the grave-acute distinction. Some of spectrum showed on the boundary of two category or opposite property from listener's identification. There were some utterances provide not clear and unambiguous indication of place of articulation. For these utterances, the direction of formant motions, or the stimulus duration, were apparently used by the listeners to resolve the ambiguity with regard to place of articulation.

Probably, interaction among the burst spectrum, the onset formant frequencies, and the direction of formant transition has to be considered to overcome 20% ambiguity. Optimized feature vector and multi-staged decision by repeated discriminant analysis between two classes of consonants in turn so as to find more effective parameter set may also rise score.

Summary

We determined the extent to which the onset spectra of natural CV utterances consisting of [p, t, k] in the environment of five vowels could be correctly classified by the linear discriminant functions. Over 80% of the utterances were correctly classified independent of vowel context and speaker difference. Thus, the spectrum at the burst onset does seem to produce, also in Japanese, an invariant gross shape for place of articulation which Blumstein and Stevens (1979) have claimed.

References

Blumstein, S. E., and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. 66, 1001–1017.

Blumstein, S. E., and Stevens, K. N. (1980). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. 67, 648–662.

Blumstein, S. E., Isaacs, E., and Mertus, J. (1982). "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. 72, 43–50.

Dixon, W. J., and Brown, M. B. (1977). *BMDP-77*, University of California Press.

Doshita, S. (1965). "Studies on the analysis and recognition of Japanese speech sounds," Thesis, Kyoto University.

Fant, G. (1960). *Acoustic Theory of Speech Production*, (Mouton, The Hague).

Halle, M., Hughes, G. W., and Radley, J. P. A. (1957). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. 29, 107–116.

Ide, K., Homma, S., Makino, S., and Kido, K. (1982). "Consonant recognition using time spectrum pattern," Trans. Comm. Speech Research, Acoust. Soc. Japan, S82–23.

Kobatake, H., and Noso, K. (1980). "Feature extraction and recognition of Japanese voiceless stop consonants by principal components analysis," J. Acoust. Soc. Japan(E), 1, 215–228.

Lachenbruch, P. A. (1975). *Discriminant Analysis*, Macmillan Publishing Co.

Lahiri, A., and Blumstein, S. E. (1981). "A reconsideration of invariance for place of articulation in stop consonants: Evidence from cross-language studies," J. Acoust. Soc. Am. 70, S39.

Markel, J. D., and Gray Jr., A. H. (1976). *Linear Prediction of Speech*, Springer-Verlag.

Mikami, N., and Ohba, R. (1981). "Three-pole linear prediction method of extracting feature parameters for the voiceless stops classification," Trans. IECEJ. J64–A, 1000–1006.

Mines, M. A., Hansen, B. F., and Shoup, J. E. (1978). "Frequency of occurrence of phonemes in conversational English," Lang. Speech 21, 221–241.

Searle, C. L., Jacobson, J. Z., and Rayment, S. G. (1979). "Stop consonant discrimination based on human audition," J. Acoust. Soc. Am. 65, 799–809.

Tanaka, K. (1981). "A parametric representation and a clustering method for phoneme recognition—Application to stops in a CV environment," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-27, 1117–1127.

(Aug. 31, 1982, received)