STUDIA PHONOLOGICA XXI (1987)

Discriminant Analysis of French Stops and Nasals

Shigeyoshi KITAZAWA

ABSTRACT

Studies on the invariant features of Japanese stop and nasal consonants have been extended to French. Place and or manner of articulation of 7 consonants /ʔ, p, t, k, b, d, g/ are discriminated in an environment of /a, o, œ, e, ɛ, u, y, i, ã, ɛ̃, õ/. The feature vector is 23 LPC cepstrum coefficients at every 10 ms of the initial 100 ms (30 ms before the burst and 70 ms after the burst). The burst point was manually specified referring to waveform display. Three of nasal consonants /m, n, ɲ/ are discriminated in the same environment. The release point, openning of the oral passage, was manually specified also. The stepwise discriminant analysis in the SAS system obtained a reduced feature set and discriminant score. The 40 male speakers' sample comprizes 3080 stops and 1120 nasals. Speaker and vowel independently stops are better discriminated than Japanese, but nasals are equivalent or a little lower than Japanese. The conclusion that the spectral pattern near the stop burst is a good feature for place discrimination can be generalized throughout French and Japanese. Results are compared with perception test.

1. INTRODUCTION

Phoneme is a linguistic concept, however, technically continuous speech is segmented as a sequence of phonemes. The problem of continuous speech recognition may be approached by precise phoneme recognition. Phoneme is variable depending on context, speaker and language. Consonants are studied usually with context, i.e. transition to/from adjacent vowel. Current speech recognition systems are very well tuned for a specific speaker but have to be adapted for different speakers. Similar phonemes in different language are characterized such as aspirated and nonaspirated stops.

But each consonant has essential articulatory movement, for example, bilabial stop starts with the closing of lips among other articulators and they apart suddenly. Similarily dental and velar stops use specific articulator and articulatory points. The similar articulation point is used also for nasals. Since the articulation is similar even for different contexts, speakers and probably languages, consonants must share

Shigeyoshi KITAZAWA (北澤茂良): Associate Professor, Department of Computer Science, Shizuoka University, Johoku 3–5–1, Hamamatsu 432, Japan.
The author has done the research under the direction of Dr. Shuji Doshita, Professor of Information Science, Kyoto University.

common property in the acoustic domain.

Consequently invariant features for consonants are hypothesized. But how and what features can be extracted? The stop burst spectrum is the most possible feature known. There are researches or experts who have shown interesting features. But intrinsically obtained features of spectrum do not seem to be generalized to unknown speakers and to different languages if they are deduced from small number of samples or speakers. Number of exceptional cases have to be integrated into a general rule which is a very difficult task.

Instead of this deductive feature extraction, one can semi-optimally reach to invariant feature by statistical analysis. Once collecting sufficient number of observations, statistical procedure can find automatically a solution for that, and the solution is a set of possible features. So the only necessary thing to do is to find what kind of observation is suitable for obtaining invariant features. In the most possible extent, observations deviate as little as possible, and deviation reflects only the phoneme difference.

Based on these hypothesis the author has been studying Japanese stop consonants and their invariant features for the place of articulation [1, 2]. He discriminated place and/or manner of articulation of the 7 consonants (/ʔ, p, t, k, b, d, g/) in an environment of 5 vowels (/a, i, u, e, o/). Among experiments the best results have been obtained observing critical band spectrum of the 70 ms after stop burst point and the 30 ms before. Nearly 90% correct discrimination was possible among stops of the same manner (/ʔ, p, t, k/ and /b, d, g/). Similarly concerning to nasals, the discrimination rate among /m, n, ŋ/ was 86% [3].

The similar method was tried to extend to French. We will also show linguistic dependency.

2. Processing Procedure

Processing starts with analysis point specification, followed by acoustic analysis and statistical analysis.

The burst point was determined as precisely as possible for all stop syllables from visual examination of waveforms. The burst conveys the most information of stop consonant and it is possible to be determined uniquely for each utterance. For vowels, vowel onset is the possible point to detect acoustically. For some vowels, burst like articulatory noise appears before vowel onset, however, point detection should be knowledgeless, therefore such noise is detected as a burst. For voiceless stops, the burst point is easily determined. The burst of a few bilabial stops, since French stops are unaspirated, is very weak. Voiced stops start with prevoicing murmur in most cases, but the burst after the onset of murmur is detected. In some cases, the burst is too weak or there is no burst. Even in such cases, the most likely point was uniquely decided.

The corresponding analysis point of nasal consonants is the oral passage open-

ing point following to the nasal murmur. This point, the release point, is basically the same point as used to analyse Japanese nasal consonants. We experienced significant difference in nasal murmur between French and Japanese.

The characteristics of French nasal speech can be described as follows: The acoustic energy of nasal murmur is larger than that of Japanese, while the acoustic energy of vowel is reduced because of nasalization of the following vowel. Therefore, it is very often difficult to determine the release point or the boundary between nasal murmur and the following vowel from the envelope of speech waveform. Global pattern of waveform envelope seems not useful for French unlike for Japanese in determining the release point. The local pattern in the waveform, however, was useful, for example, which supposedly corresponds to a small the phase change at the release point. This feature alone does not define uniquely the release point, however, can be used in reference with waveform envelope, zero-crossing and others

The acoustic processing and statistical analysis are similar between stops and nasals. Based on the burst or the release point detected visually, time varying spectrum was observed. Acoustic parameters are 23 LPC cepstrum coefficients. The burst or the release spectrum is the most important feature of consonant place of articulation. Besides prevoicing murmur spectrum and transitory spectrum is also observed. Hamming window of 256 sampling points at 16 kHz (effectively 15 ms) was shifted every 5 ms, averaged in three frames overlapping one frame, consequently spectral information was obtained every 10 ms.

Discriminant experiment was performed on the spectral pattern obtained from acoustic analysis. First, features are selected for reduction of dimensionality. By apriori knowledge, too fine spectral structure is not important but noisy deviation. So in cepstrum dimension, lower components are useful but higher components may be useless. In time domain, spectral change at the burst point is large but in transitory part the spectral change is gradual and the spectral pattern does not change much between two adjacent frames. Therefore some cepstrum coefficients and frames are redundant. Then, assuming the equality of covariance matrix, linear discriminant function was used. These statistical analysis procedures STEPDISC and DISCRIM are employed from the well designed software SAS (Statistical Analysis System).

3. SPEECH DATA BASE

Syllables are chosen as a most simple context. We intended to include as many vowels in French as possible while keeping the amount of data set within reasonable size and also had in mind the easyness of pronounciation for common people. In French there are 3 voiceless stops and 3 voiced stops, bilabial, dental and velar. Among 16 vowels in French, some are difficult to distinguish for people, so we selected 11 vowels which cover almost all. Isolated vowels are included as

a null consonant assigned a phonetic symbol /ʔ/ (glottal stop). Consonants /ʔ, p, t, k, b, d, g/ combined with vowel /a, o, œ, e, ɛ, u, y, i, ã, ɛ̃, õ/ composed 77 different syllables. Among nasal consonants we examined three of them /m, n, ɲ/ in combination with 11 vowels, except that /ɲ/ combined with 6 back vowels, composed 28 different syllables.

These syllables are pronounced once by each speaker. Speakers included are 40 native French or fluent French speaking males. Total syllables processed are 3080 of stops and 1120 of nasals. Speech was recorded in a quiet studio at ENST using a dynamic microphone and PCM processor according to GRECO standard procedure, then digitized at CNET into 16 bit 16 kHz standard 1600 bpi tapes.

4. Results

Recognition performance were evaluated by minimum distance in terms of generalized square distance. The correct recognition rate was compared for within data set experiment and unknown speaker experiment. In the within dataset experiment, the covariance matrix was computed from all the available data, and using this matrix the generalized square distance was computed and error rate was evaluated. In the unknown speaker experiment, samples from one speaker is reserved as a test data, the rest of samples are used to compute discriminant functions. Error rate was observed concerning the one reserved speaker. The speaker changed in turn and errors were averaged.

As a result of stepwise selection of discriminant variables (ten frames of 0th to 23rd cepstrum coefficients), frames 1, 2, 3, 4, 5, 6, 8, 10 are included, but 7 and 9 are omitted because they are redundant, and cepstrum parameters higher than 17th are also omitted. Finally, 77 cepstrum coefficients are selected as significant ($F>4.0$) and included for further discriminant analysis. The most significant frame in terms of number of coefficients included is 3, i.e. the burst spectrum, and lower coefficients in the 1st frame are also significant because of prevoicing distinction between voiced and voiceless. Even the 10th frame is still contributing to discrimination.

The recognition score of within datset experiment is quite high as 90%. Confusion from voiceless to voiced is zero, from voiced to voiceless is a few except 8 confusions from /b/ to /p/ and 9 confusions from /g/ to /k/. Confusion from /t/ to /k/ is the largest, next /ʔ/ to /p/ and /k/ to /t/. Among errors in voiced stops, confusion from /b/ to /d/ is significant. Table 1 shows the result of unknown speaker experiment. Average recognition rate is 87%. Drop of 3% compared to within dataset experiment means statistically stable result, i.e. sufficient number of samples are observed. However some errors between phonemes increase from 30 to 48 (/ʔ/ to /p/). This means the number of samples is insufficient to estimate errors between individual phoneme pairs.

The number of errors of individual speakers deviated from 2 to 25 among 77

Taebl 1. Stop consonant recognition rate for unknown speaker (one speaker is left for evaluation from 40 speakers and the discriminant function was designed from 39 speakers)

	classified into						
from	ʔ	p	t	k	b	d	g
ʔ	.81	.11	.03	.04	.0	.0	.0
p	.08	.84	.04	.04	.0	.0	.0
t	.0	.05	.85	.09	.0	.0	.0
k	.0	.02	.08	.90	.0	.0	.0
b	.0	.02	.0	.0	.86	.06	.05
d	.0	.0	.2	.0	.03	.92	.03
g	.0	.0	.0	.03	.03	.04	.90

total recognition rate is 87%

phonemes examined. For the worst speaker, bilabial stop burst is too weak to be detected and 10 voiced stops are difficult to determine the burst point. For the next worst speaker, aspirations are not seen in all cases. The frequent errors are between /t/ and /k/ before /i/ or /y/ and between /p/ and /ʔ/ before /ɛ/.

Exactly the same procedure as for stop consonants was used also for nasals. By stepwise discriminant analysis, 1st, 2nd, 3rd, 5th, 8th and 10th frames are chosen among 10 frames. Other frames, 4th, 6th, 7th and 9th frames, are discarded. Contrast with stops, the speed of spectral change is slower in nasals than stops. Finally 54 variables are chosen among these frames.

Nasals differ from stops in the range of effective cepstral coefficients; the higher order cepstral coefficients are effective. Variables are selected among 24 cepstral coefficients in each frame. The 5th frame supplied the largest number of variables. This frame coincide with the onset of the following vowel. Corresponding nasal murmur frames, 1st, 2nd and 3rd, bring large number of variables. Number of variables are chosen even from the 10th frame which correspond to stationary vowel. Those variables from 5th, 8th and 10th frames are supposed to represent transitory information. A few variables are chosen from the 8th frame, this means that spectral change at this point is slow and this part of change can be interporated

Table 2. Nasal consonant recognition rate for unknown speaker (one speaker is left for evaluation from 40 speakers and the discriminant function was designed from 39 speakers).

	classified into		
from	m	n	ɲ
m	.86	.09	.05
n	.10	.83	.07
ɲ	.05	.08	.86

total recognition rate is 85%.

from the beginning and ending points of the glide spectra. At the beginning of glide, the 5th frame, provided higher order cepstrum coefficients, on the other hand, at the end of glide, 10th frame, provided lower cepstrum coefficients.

The resultant nasal confusion matrix of unknown speaker is shown in Table 2. The discriminant score was 89% in apparent, but the average of unknown speaker droped 4% to 85%. Difference in scores between individual speaker valies 0 to 12 errors among 28 syllables. Two speakers' nasals are perfectly recognized. Those speeches in which the boundary between nasal murmur and vowel onset can be unambiguously determined are well recognized. Among errors between syllables, the error from "ni" to "gni" was the most frequent, however, this can be explained that some speaker pronounce a syllable "ni" as phoneme /n/ but some /ɲ/. Sometimes these two are confused in this context.

5. Discussion

There are several alternatives of point observations, a starting point of prevoicing of voiced stops, a voice onset point at the beginning of vowel formant transition, and the ending point of formant transition. The release point preceeds the vowel onset, which is analogous to the voice onset time defined for stop consonant.

Comparison with results for Japanese, French stops are much better recognized, though the number of vowels included for French is 11 but 5 for Japanese. Stop consonants seem to be vowel independent. The significant difference is much less errors between voiced/voiceless in French. This is due to the fact that most French voiced stops have prevoicing vibration before burst but large number of Japanese have not. This is a kind of linguistic dependency.

Nasals are not as easy as stops to be recognized by machine. Discriminant rate of French nasals is lower than Japanese, however, the conclusion is not yet evident. The analysis procedure was not complete in several ways; especially in accuracy of the release point. If the release points were determined more precisely, the recognition rate would be surely augmented. In order to obtain accurate release point, the more precise acoustic analysis is necessary; for example, high pass filtering, inverse filtering using linear prediction parameter estimated during nasal murmur segment, spectrogram, fundamental frequency change, and formant frequency values. Not only frequency domain feature, but also time domain feature such as phase information will be helpful.

In order to compare the recognition results by machine with human perception, a perception experiment was performed on the syllables in 5 speakers' among all speakers. Randomly chosen 200 syllables are presented to 11 subjects who are asked to answer the syllable they heard. The experiment was done in a quiet room using audio-cassette taperecorder and speaker system with comfortable loudness level. Each syllable was presented once every 2 second interval. Table 3 shows confusion matrix of human perception.

Table 3. Confusion matrix of perception test.

from	perception ?	p	t	k	b	d	g	m	n
?	251	1							
p	14	215	1		1				
t			175	1					
k				252			1		
b		2	1		265	6			
d			1			339	12		
g				1			175		
m								209	11
n	1					2		12	249

The result shows that human can discriminate stop consonants very well without linguistic semantics. The tendency of errors coincides with that of machine, for example, consonant /p/ drops often, /d/ is often confused with /g/. The author himself participated in this experiment and could recognize stop consonants quite well. As far as this experiment is concerned, French stop consonants can be recognized by a Japanese as good as native speaker. The feature of stop consonant is universal.

Only /m/ and /n/ are perceptually examined among nasals. This shows that even for human it is difficult to be perfect in nasal syllable recognition but about 5% error is unavoidable. Nasals are more difficult to recognize than stops.

Vowels are also asked to answer in the experiment as well as consonants. Native speakers could recognize vowels 98% correctly almost perfectly. On the other hand, vowel recognition was difficult for foreigners. Discriminant boundary of vowel categories is language dependent and difficult to adapt.

6. CONCLUSIONS

French stop consonants are shown through statistical analysis to be recognizable by machine using features of spectrum adjacent of the stop burst under speaker and vowel independent condition. The similar experiment with nasal consonants are shown to be possible but a little difficult. These results are confirmed through the perception experiment of example syllables.

Since the equivalent results were obtained on French as we have got on Japanese before, the hypothesis that the spectrum at the burst of stops and at the release of nasals is one of physical form of the invariant feature of consonant was generalized throughout at least two languages.

ACKNOWLEDGEMENT

This work has been done while the author has been staying at ENST as an exchange scientist of CNRS-JSPS cooperation. The author would like to express

special thanks to Professor J. P. Tubach who organized recording and perception experiments, and to all people who kindly helped me.

References

1) S. Kitazawa and S. Doshita, *Proc. of Seventh International Conference on Pattern Recognition* (Montreal, Canada, July 30–August 2, 1984), 179–181.
2) S. Kitazawa and S. Doshita, *Proc. of IEEE-IECEJ-ASJ International Conference on Acoustics, Speech, and Signal Processing*" (Tokyo, Japan, April 7–11, 1986), 2703–2706.
3) S. Kitazawa and S. Doshita, *Proc. of The Eighths International Conference on Pattern Recognition* (Paris, France, October 27–31, 1986), 48–50.

(Aug. 31, 1987, received)