



STATISTICAL DISCRIMINATION OF FRENCH INITIAL STOPS

S. KITAZAWA*, J.P.TUBACH**

ABSTRACT

Studies on the invariant features of Japanese stop consonants have been extended to French. Place and or manner of articulation of 7 consonants /ʃ,p,t,k,b,d,g/ are discriminated in an environment of /a,o,œ,e,ɛ,u,y,i,â,ê,ô/. The feature vector is 23 LPC cepstrum coefficients at every 10ms of the initial 100ms (30ms before the burst and 70ms after the burst). The burst point was manually specified referring to waveform display. The stepwise discriminant analysis in the SAS system was used to obtain reduced feature set and discriminant score. The sample comprises 3080 monosyllables from 40 male speakers. Speakers and vowel independent discrimination results better than for Japanese stops. 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.

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. Similarly dental and velar stops use specific articulator and articulatory points. Since the articulation is similar even for different contexts, speakers and probably languages, consonants must share 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 spectrogram reading 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 numbers 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 dependancy on the phoneme type. The burst point is suitable as the reference point of measurement, because it can be determined uniquely and reliably for each syllable and the spectrum extracted from this point is characteristic of the phoneme.

* ENST Dept. SYC. (CNRS, U.A.820), and also with Shizuoka University, Johoku 3-5-1, Hamamatsu 432, Japan ** ENST Dept. SYC. (CNRS, U.A.820), 46 rue Barrault, 75634 PARIS cedex 13, France

Based on these hypothesis the first author(S.K.) 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 70ms after stop burst point and the 30ms before. Nearly 90% correct discrimination was possible among stops of the same manner (/ʔ,p,t,k/ and /b,d,g/). The similar method was tried to extend to French. We will also show linguistic dependency.

PROCESSING PROCEDURE

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

The burst point was determined as precisely as possible for all 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, so 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 launch 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.

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

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).

SPEECH DATA BASE

Syllables are chosen as a most simple context. We intended to include as much vowels in French as possible while keeping the amount of data set within reasonable size and also had in mind the easyness of pronunciation 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,ɛ,õ/ composed 77 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. 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 16bit 16kHz standard 1600bpi tape.

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 distances was computed and error rate was evaluated. In the unknown speaker experiment, sample from one speaker is reserved as 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. burst spectrum, and lower coefficients in the 1st frame are significant because of prevoicing distinction between voiced and voiceless. Even the 10th frame is still contributing to discrimination.

The recognition score of within dataset experiment is quite high as 90%. Confusion from voiceless to voiced is zero, from voiced to voiceless is few except 8 confusions from /b/ to /p/ and 9 confusions from /g/ to /k/. Confusion from /t/ to /k/ is 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 simulated 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 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 /ε/.

Table 1. Stop consonant recognition rate for unknown speaker (one speaker is left for evaluation from 40 speakers 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%

DISCUSSION

There are several alternatives of point observations, a starting point of prevoicing of voiced stops, a voice onset point at vowel formant beginning, and an ending point of formant transition. Dynamic time warping may also be effective.

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 quite few 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.

Perception test was conducted to support the recognition experiment. Just 5 speaker's syllable are randomly selected to compose 156 test syllables are presented to 11 listeners to recognize. Table 2 shows a confusion matrix concerning the consonant identification. Human being is a quite good recognizer of stop consonant but there are correlations with errors by machine. Consonant /p/ tends to be missed, and /d/ is often confused with /g/.

CONCLUSIONS

French stop consonants can be discriminated very well by statistical analysis of burst spectrum independent of speaker and vowel. Conclusion that spectral pattern near the stop burst is speaker and vowel independent feature for stop place discrimination is general throughout French and Japanese.

ACKNOWLEDGEMENT

This work was done while the first author was staying at ENST as an exchange scientist of CNRS-JSPS cooperation.

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.

Table 2 . Confusion matrix of perception test.

	perception						
from	?	p	t	k	b	d	g
?	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