PERCEPTUAL EVALUATION OF SPECTRALLY CONFUSING STOPS AND NASALS

Shigeyoshi Kitazawa
Shizuoka University, Hamamatsu, JAPAN.

ABSTRACT

Spectral similarity does not necessarily correlate to perceptual similarity. Bayesian classifier showed overlaps between consonants in spectral representation. The perceptual test of misclassified consonants was 90 % correct. Concerning 10 % of low intelligible consonants, half of the misperceptions corresponded to spectral deviations. The remaining misperceptions showed a systematic tendency which can be interpreted in terms of distinctive features.

1. INTRODUCTION

We intended to certificate the intelligibility of speech data used for speech recognition by machine. Since observed recognition errors may be due to low intelligible speech, we designed a perception test of stop consonants misclassified with a Bayesian classifier. This experiment unintentionally correlated to the hypothesis that speech sounds are perceptually decomposed into distinctive features.

This is to study the difference between machine and human being with respect to the recognition constituents. The reason why a machine does not distinguish speech sounds which a human can easily hear is discussed through the perception test of natural speech.

2. PROCEDURE

The context we have chosen was simple syllables consisting of a consonant and a vowel following. The language is French. The phonemes tested include 10 consonants /?,p,t,k,b,d,g,m,n,gn/ followed by 11 of 16 French vowels /a,o,eu,e,ai,ou,u,i,an,in,on/. We assumed a phoneme /?/ before the isolated vowel syllables. All of the syllables came from 40 male speakers living in Paris. The test sample consists of 4200 independent phonation of syllables. Most speakers are native French. In order to qualify articulation of the speech data base, 200 syllables selected randomly from 5 speakers' speech were presented to 11 listeners in a quiet listening room. French speaking listeners could identify consonants with more than 97 % of accuracy.

3. STATISTICS

As a certificate of the speech quality, syllables were classified according to the initial consonant[1]. The acoustic parameters were 28 LPC cepstrum coefficients using a 256 point Hamming window (effectively about 15 ms width under 16 kHz sampling frequency) shifted every 5 ms. Along these windows, the cepstrum coefficients are averaged over 3 consecutive frames resulting a set of 10 frames of smoothed cepstrum coefficients at every 10 ms.

The burst point of stop consonants and the release point (opening of the oral passage) of nasal consonants are determined as precisely as possible through visual inspection of waveforms. For initial vowels, initiation of vibration was inspected. The third analysis frame is at this critical point determined. With these procedure, consonant specific features are extracted.

The classification is based on the multi-dimensional statistical analysis of the above shown 290 scalars for each sample. The stepwise discriminant analysis selected around 50 to 70 elements of the vector. Then, Bayesian classifier determined the correct classification rates . In the separate analysis, 87 % for stops and 85 % for nasals were the scores.

Syllables tested were 4200 from 40 male speakers and 425 misclassifications were observed as shown in Table 1.

4. SPEECH PERCEPTION

Inspection of previous reports and our own experience show that syllables are highly intelligible if heard under low noise and wide frequency band condition. In our experimental paradigm, those syllables in which consonants are correctly classified are regarded to be intelligible, and those misclassified are subjects for perceptual experiments. Our preliminary experiment showed that most of the syllables were highly identifiable (97 % in average).

Syllables used for perceptual tests were 425 which were misclassified in the closed discriminant analysis[2]. This list contains all the possible syllables except /ou/, /teu/, /kon/. Each syllable were recorded on an audio cassette tape at a sampling rate of 16 kHz. The listeners heard the stimuli through headset, one syllable each 4 sec, and identified the syllables by writing.

All the 11 listening subjects are native French speakers. Records were kept of all responses.

5. RESULTS

Effective responses were 4620, among which 655 misperception were observed, therefore 86 % was correctly perceived. Half of the correct answers were unanimous among all listeners. The first finding means imperfection of recognizer, that is we missed some important features of consonants but puzzled with phantom features.

Among misperceptions, 462 are concerned with consonants(Table 2.), 237 of them coincide with machine errors, and the rest 225 were different perception from machine errors(Table 3.). Misperceptions concerning vowels was 231, which consists of 193 vowel errors and 38 consonant and vowel errors. Table 2. and Table 3. show in percent of each consonant presentation. Subtraction of Table 3 from Table 2 gives coinciding errors.

About half of consonant misperceptions coincided with misclassifications. This means acoustic features used for classification reflect perceptual similarities. Amongst all, 7 syllables are coincidentally misperceived by 10 of 11 hearers. Inspection of the waveforms showed that 3 errors from /b/ to /p/ and one /d/ to /t/ were not proceeded by prevoicing. In /bu/ to /u/ case, both prevoicing and burst were not observable. In /t/ to /p/ transition, very fast rising of amplitude at the onset without fricative noise was clearly observed.

The average correct response rate was 90 % which is 7 % lower than average. As usual, errors tended to accompany specific vowels or to concentrate to specific speakers and listeners. The score deviated from 87 % to 95 % between listeners. Five of the speakers also participated in the listening test. They can hear their own voice better than others.

Observing confusion matrices, we can find characteristic distributions. The perceptual confusions, Table 3, distributed along the diagonal and dencer in the upper triangular matrix. On the other hand, the matrix of machine recognition, Table 1, distributed differently. This is shown more clearly in Table 2 as the machine specific error distribution. Deviation to the lower triangular matrix is very significant comparing to the almost equal distribution in machine error (Table 1.). Another comparison with Bayesian classifier is in Table 4 as human specific perceptions. The distributions are a litte sparse to draw definite knowledge, however, rather frequent in the upper triangular matrix.

In these asymmetry of matrices, there is some specific characteristics of human perception. Confusions observed in Table 3 were sorted in terms of distinctive features as in Table 5. The table indicated meaningful tendency of perceptual transition. We will discuss in the following section.

6. DISCUSSION

Perception test of misclassified consonants is a unique experiment where several factors are combined; insufficiency of the features used by a classifier, difference of the perceptual space of speaker and hearer etc.. Since speakers hear their own voice, speaker recognize their speech to be correct. Definitely most of speeches convey sufficient acoustic information. Normally, the error rates of these speeches are very low, so it would take a long time to obtain accurate estimates of the error probabilities. However, misclassified consonants are low intelligible syllables or low intelligibility items which cause significantly higher error rates.

The importance of distinctive features in perception of consonants was demonstrated. For each feature, one feature specification (+ or -) tended to dominate over the other. As demonstrated in Tables 2, 4 and 5, there was

for each feature an asymmetry in the frequency of + and - feature specifications in error responses. With the exception of anterior, the dominant feature specifications are all "unmarked", according to traditional phonological theory. One plausible explanation for the dominance of unmarked feature specifications is that the low intelligibility of selected syllables leads to a simplification of the percept (i.e., a loss of information). In some of the perceptual shifts, acoustic features such as loss of prevoicing and weakened burst noise were observable.

The results from the present experiment are highly compatible with those from previous studies.

Previous paradigms include proximity estimates[3], identification of masked or distorted speech[4], dichotic presentation[5], recall test with the short term memory[6], and natives vs. non-natives[7]. However, much of this research dealt with listening conditions acoustically degraded or loaded stresses on listeners. Such researche has provided ample evidence that the number of *distinctive features* play an important role in perception of consonants and that the phonemes are not a perceptual unit. On the other hand, phonemes are a unit of classification.

The proximity estimates assume symmetry of the distance matrix. The analysises of MN test data also assumes symmetry of the confusion matrix. On the other hand, dichotic listening and short term recall tests are substantially asymmetric. Wickelgren did not mention about asymmetry of confusions or tendencies observed in distinctive feature system. Hayden explicitly indicated the feature specification dominance and suggested the perceptual system to favor the simpler (unmarked) feature specification in the presence of competing cues.

7. CONCLUSIONS

The purpose of this study was to reveal that the simple acoustic comparison is insufficient to explain perceptual differences of consonants. Human listeners can show essentially higher performance than machines but have different characteristics. The speaker independent acoustic analysis showed more than 90 % correct discrimination between consonant place of articulations. Those syllables misclassified by Bayesian recognizer are further examined. These outliers are an interesting set of examples providing an insight into human perception and production of speech. Most of them are phonetically perfect but uncovered by recognizers and a few of them are imperfect productions.

Perceptual experiments, using native listeners, exhibited a high intelligibility except for some acoustically confusing syllables. We found listeners made confusions under natural hearing condition. Half of the incorrect answers coincided with the misclassifications of the recognizer, perhaps through the similar evaluation of the features. Asymmetric distribution of the confusion matrix suggested that there are differences in strategy between human and machine.

The last point is important in relation to the hypothesis that speech sounds are perceptually decomposed into distinctive features. Analysis showed the tendency that perceptual system favors the simpler (unmarked) features in the presence of low intelligible cues. On the other hand, recognizers minimize the total errors by distributing errors among possible solutions.

The findings suggest that distinctive features play an important role for human perception of phonemes.

ACKNOWLEDGEMENTS

This work is an extension of cooperative work with Professor J.P.Tubach at ENST, Paris. People at ENST and ATR kindly participated in our perception tests. We appreciate conveniences offered from ENST and ATR.

REFERENCES

[1] KITAZAWA,S and J.P.TUBACH (1987), "Statistical discrimination of French initial stops," *Proc. European Conf. Speech Tech.*, 1, 91-94.

[2] KITAZAWA,S and J.P.TUBACH (1988), "Discriminant analysis and perceptual test of French stops and nasals," *Proc. 9tn Int. Conf. Pattern Recognition*, 1077-1079.

[3] BLACK,J.W.(1970), "Interconsonantal differences," *Essays in Honor of Claude M. Wise*, (Brownstein,A.J. et al. edt.) Artcraft Press,' Missouri, 74-96.

[4] MILLER,G.A. and NICELY, P. E. (1955), "An analysis of perceptual confusions among some English consonants," *J.A.S.A.*, 27, 338-352.

[5] HYDEN,M.E. and KIRSTEIN,E.

and SIGH,S.(1979),"Role of distinctive features in dichotic perception of 21 English consonants," *J.A.S.A.*, 65, 1039-1046.
[6] WICKELGREN,W.A. (1965), "Distinctive features and errors in short-term memory for English consonants," *J.A.S.A.*, 39, 388-398.
[7] SIGH,S. and BLACK, J. W. (1966), "Study of twenty-six intervocalic consonants as spoken and recognized by four language groups," *J.A.S.A.*, 39, 372-387.

Table 1. Machine Recognition Rate for Unknown Speakers

Classified	Actual Consonant									
	[ʔ]	[p]	[t]	[k]	[b]	[d]	[g]	[m]	[n]	[ɲ]
[ʔ]	.81	.08	.0	.0	.0	.0	.0			
[p]	.11	.84	.05	.02	.02	.0	.0			
[t]	.03	.04	.85	.08	.0	.2	.0			
[k]	.04	.04	.09	.90	.0	.0	.03	Not Examined		
[b]	.0	.0	.0	.0	.86	.03	.03			
[d]	.0	.0	.0	.0	.06	.92	.04			
[g]	.0	.0	.0	.0	.05	.03	.90			
[m]	Not Examined							.86	.10	.05
[n]								.09	.83	.08
[ɲ]								.05	.07	.86

Table 2. Intelligible Bayesian errors.

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ		7	2	3			6			
p	34			3	6					
t	12	13		40		4	3			
k	17	5	40		2		9			
b						15	3			
d					32		14			
g					30	27				
m									26	35
n								43		26
ɲ								14	28	

Table 3. Perceptual Confusions

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ	99	10	2	2	2		.3		.2	
p	.3	86	11	5	6	1		.2		
t		3	84	2		7				
k		1	1	88			1			
b		.8	.2		83	1	7		.2	
d			1		4	88	4			
g			.2	3	1	3	86			
m					4			98	5	
n								1	93	3
ɲ				.3			.8	.4	2	97
etc.	1		.3	.6	.4	.4				1

Table 4. Perceptions off the Bayesian errors.

obs.	*real consonants*									
%	ʔ	p	t	k	b	d	g	m	n	ɲ
ʔ		3	2	2	.6		.3		.2	
p	.2		4	4	4	1		.2		
t		1		.3		4				
k		.2	.5							
b		.8	.2			.4	2		.2	
d			1		.4		.8			
g			.2	3	1	1				
m					4				.3	
n								1		
ɲ				.3			.8		1	
etc.	1		.3	.6	.4	.4				.4

Table 5. Percentages of feature specification for perceptual errors.

features	*%+specification*	*%-specification*
Coronal	8.52	17.72
Anterior	16.16	5.06
Voiced	5.01	13.76
Consonantal	0.32	15.76