DP-MATCHING: WITH OR WITHOUT PHONEMES?

Shigeyoshi Kitazawa*, Masa-aki Ishikawa** and Shuji Doshita**

*   Shizuoka University,Hamamatsu-shi,432,Japan
** Kyoto University, Sakyo-ku, Kyoto-shi,606,Japan

ABSTRACT

Attempts at automatic speech recognition have known several waves.

Early efforts were based on the faith that speech is a string of phonemes that can be isolated and recognized one by one. This wave broke when it became clear that the physical realization of a phoneme is smeared in time and mingled with that of its neighbors, and also context and speaker-dependent.

Next came the invention of the highly successful time-warping DP-matching methods, in which whole words are matched by templates. This wave is still going strong, at least in Japan, but it may have reached a high mark.

To probe this question, we investigate the case of the "jion'', a subset of character readings that "generates" a large subset of Japanese. This set has low redundancy and contains many minimal pairs. Error analysis of DP-matching shows that most errors occur between pairs that differ only in their initial consonant, especially if it belongs to groups such as plosives or nasals.

Combining DP-matching with limited-scope phoneme recognition could break through present limits.

I INTRODUCTION

When we listen to speech in an analytic frame of mind, we hear it, or we think we hear it, as a succession of phonemes. It seems to us that we could pick out each "phoneme" if only they didn't flow past quite so quickly. This view now seems naive (Repp 1981), but it was natural enough when speech recognition began.

Systems based on phoneme recognition run into a variety of troubles. First, the boundary between successive "phonemes" is elusive (the segmentation problem). Second, once a phoneme segment is fenced off, it is found that it bears little ressemblance to the same phoneme uttered elsewhere in the speech stream, or in isolation (the co-articulation problem). Worse still, the range of possible realizations may overlap that of a different phoneme. Finally, even if a taxonomy of all phonemes in context is attained, it proves different from speaker to speaker (the speaker-dependency problem).

The task of designing systems to reliably extract and sort out all "phoneme" cases and cues is thus formidable. It drained the energy of early researchers, and the results were disappointing.

The invention of Dynamic Programming time-warp matching came as a relief because it provided a simple, elegant, and immediately appliable method of recognizing whole words. Many variants of DP-matching have been proposed (continuous, multiple level, augmented, etc.) and its efficiency has been improved to cope with large vocabulary, multiple speakers, etc..

DP-matching continues to be the object of much research in Japan (27 papers out of 200 on speech at the 1984 meetings of the ASJ). However, it may be that the efficiency of DP-matching has reached its maximum. The fundamental drawback is that the discriminating distance is calculated over a whole word. If two words differ by just one phoneme (minimal pair), the difference is "diluted" and may be masked by small variations over the rest of the word.

This is particularly so for "short" phonemes (eg plosives) in long words. These are likely to cause problems if they occur in an application's word list.

DP-matching scores are often evaluated on lists of city names. The results cannot easily be extrapolated because of the inhomogeneity and redundancy of cues in such sets. For this reason, we chose instead to perform our experiments on a set of words, the "jion", that are highly representative and have low redundancy.

The word "jion" means "character sound", and designates the sounds that the "Chinese" readings of Sino-Japanese characters can assume. Many words in Japanese are built up of "jion", so combinations of "jion" cover most of the language. Many "jion" are minimal pairs, and these are representative of longer minimal pairs in which they occur.

The "jion" set is thus a good evaluation set, and we used it to try to situate the limits of DP-matching recognition methods. As expected, the score attained by DP-matching on the jion set is much lower than on a city name list. In addition, an analysis of the errors provided interesting results that we discuss here.

II PHONOLOGICAL STRUCTURE OF "JION"

The "jion" correspond originally to an old Chinese syllabary consisting of 403 kinds of sounds (excluding toneme difference). This syllabary was japanized and reduced when Chinese characters were introduced in Japanese.

Each "Jion" consists of two to four phonemes, forming one or two syllables. When the first phoneme is a vowel, it is assumed that it is actually preceded by a glottal stop /?/. Not all combinations of can occur, and the phonological structure is confined to the following four types: /CV/, /CVN/, /CW/ and /CVCV/, where /N/ designates the mora-nasal and /C/ is a consonant (McCawley,1968).

Table 1. Phonological structures of "jion" (ignoring initial phoneme), and number of occurences of each:

CV		CVN		CVV	
group	number	group	number	group	number
Ca	19	CaN	13	Cai	13
Ci	11	CiN	11	Cui	6
Cu	17	CuN	8	Cuu	13
Ce	10	CeN	12	Cei	12
Co	20	CoN	12	Cou	23
				Cii	1
total	77	total	56	total	68

CVCV					
group	number	group	number	group	number
Caku	23	Cuku	4	Coku	18
Catu	12	Cutu	6	Cotu	9
Cati	7	Cuti	3	Coti	6
Ciku	6	Ceku	1	Citi	6
Citu	10	Cetu	12	Ceti	8
Ciki	5	Ceki	9		
total		145			

The "jion" set has thus the following chacteristics :

a) the number of segments is limited;
b) there are only 33 kinds of phonological structures (if one excepts initial phoneme);
c) there are many minimal pairs, (for example, /baku/ and /daku/).

III DP-MATCHING OF 346 "JION"S

One male speaker produced the set of 346 "jion"s twice. We made the first set of them the templates and the second set the object of recognition. Waveforms were first low-pass filtered at 8.9 KHz and then sampled at 18.5 KHz. The parameterization, a 20th-order LPC analysis, was carried out over 20.8-ms Hamming windows shifted every 6.92-ms.

LPC cepstrum distance is used as inter-frame distance. Using these local distances, the distance between the input pattern and the reference pattern is calculated by means of a dynamic programming time warping technique. As a result 43.9% recognition rate was achieved.

Table.2 Classification of recognition errors:

type	errors	examples
initial consonant only	164 (84.5%)	a-ma, ran-nan,ta-a, mei-rei,satu-zatu, den-gen, batu-matu, bi-ri, bo-go, etc.
vowel in initial syllable	7 (3.61%)	kotu-katu,sun-son, sei-sai, sen-san, syaku-syoku, dan-don, etc.
consonant & vowel in initial syllable	14 (7.22%)	katu-hutu,kan-ton, sii-tui,siti-keti, sui-zai,seti-zati, soti-zati,nai-rui, nan-mon,ratu-botu, bati-oti, etc.
others	9 (4.64%)	so-son,syuu-syu, tyu-tyuu, hu-huu, me-men, yu-yuu, ryu-gyuu, etc.

Errors in the initial syllable account for 95.4% of all errors, of which 88.6% were errors in the initial consonant only. The recognition rate is low compared to the rate currently achieved on sets of city names, but this is precisely attributable to the low redundancy of "jion".

To illustrate this point, suppose we build a system that can recognize only the vowel parts of a word. If we input the sounds of the 346 "jion"s, the average number of symbols confused (recognized as the same word) is 19.8. If we input instead a set of 641 city names, the average number of confusions is 2.75.

Table 3 shows the recognition rates of distinctive features in the morpheme-initial consonant. The rate for "strident" was 100%, and that for "sharp" and "flat" were comparatively good. "Compact" was the worst, at 81.3%. However this result does not necessarily reflect the recognizability of the initial consonant, as there are constraints within the set that aid recognition. For example "strident" is a feature which opposes affricates /ts/ and /dz/ to simple stops /t/ and /d/. But /ts/ and /dz/ can precede only /u/, and /t/ and /d/ only a vowel among /a/, /e/ and /o/.

Table.3 Recognition rate of distinctive-features in initial consonant:

features	%	samples
1. strident	100.0	10
2. sharp	93.1	346
3. flat	90.9	11
4. consonantal	90.5	346
5. continuant	87.7	114
6. obstruent	85.7	294
7. grave	83.3	239
8. voiced	82.7	185
9. nasal	82.1	67
10. compact	81.3	80

IV CONCLUSION AND PROSPECTS FOR RECOGNITION

The "jion" set experiment showed that DP-matching scores can be rather low on a word set containing many minimal pairs. It is not a worst-case set, and one could expect performance to be even worse if the set contained more minimal pairs or longer words, as might occur in an application.

However, the experiment also showed that the errors occur in a very limited number of configurations: mainly confusions of minimal pairs differing by initial consonants belonging to the same group (for example nasals or plosives). This suggests that combining a limited-scope phoneme discrimination method with DP-matching might drastically improve recognition scores.

Over the last few years our laboratory has been working on obtaining high quality discrimination of consonants. Our first results were on plosive discrimination (Kitazawa et al 1982), and nasals (Kitazawa et al 1984).

The method used is based on statistical analysis of spectral parameters gathered over several consecutive frames. The method calculates canonical vectors that can be considered as optimal linear combinations of the parameters. The reader interested in the details should refer to the papers quoted. Discrimination results are typically 92% and 80% for plosives and nasals, respectively.

Developing similar discrimination methods for all possible distinctive features, and combining them into a phoneme recognition system would be impractical. However, by concentrating on a limited set of features, it should be possible to improve results of DP-matching.

It can be argued that speech features are designed to be recognized by humans rather than by machines. A human relies heavily on syntax, semantics and context to supplement the acoustic cues, and a machine that cannot do the same is sure to be limited in performance.

However, our opinion is that there is still much that can, and should be improved in bottom-up speech recognition before we should give up and let top-down methods take over.

ACKNOWLEDGEMENTS

Authors wish to express their thanks to Dr.Alain de Cheveigne for his criticism and suggestions.

REFERENCES

[1] Repp,B.H. "On Levels of Description in Speech Research," JASA 69, 1981, 1462-1464.

[2] McCawley,J.D. The Phonological Component of a Grammar of Japanese, The Hague,1968.

[3] Kitazawa,S. and S.Doshita, "Discriminant Analysis of Burst Spectrum for Japanese Initial Voiceless Stops", Studia Phonologica XVI, 1982,48-70.

[4] Kitazawa,S. and S.Doshita, "Speaker and Vowel Independent Recognition of CV Initial Plosives by Reduction and Integration of Running Spectra" In Proc. ICPR-84. Montreal, Canada, 1984, pp.179-181.

[5] Kitazawa,S. and S.Doshita, "Nasal Consonant Discrimination by Vowel Independent Features," Studia Phonologica XVIII,1984, 49-61.