DETECTION AND IDENTIFICATION OF PLOSIVE SOUNDS IN WORDS

MASUZO YANAGIDA YOUICHI YAMASHITA OSAMU KAKUSHO

The Institute of Scientific and Industrial Research
Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567 Japan

ABSTRACT

A system for detecting and identifying plosive sounds in Japanese words are presented. The fundamental parameter employed here to detect plosive sounds is the cepstral distance between the analysis results of an odd pair of frames with their starting positions coincided. Introduced in this report for elimination of removable candidates are short time power, pole positions obtained by low-order analyses, and a measure for representing relative disposition in the discrimination space. Phoneme identification or mutual discrimination among the detected candidates for plosive sounds is carried out by using a following-vowel dependent discrimination algorithm formerly developed by the authors.

INTRODUCTION

For continuous speech recognition of large vocabulary for unspecified speakers, discrimination by phoneme or phoneme group is required, but no reliable automatic detection/identification method has been developed for most of the phonemes, particularly for plosive sounds, whose intra-group discrimination is most difficult among phoneme groups.

For discrimination of Japanese voiceless plosives Kitazawa and Doshita[1] proposes a method using spectral information at the burst. While Tominaga et al[2] proposes discrimination of voiced plosives using transition properties form the burst to the following vowel, since Japanese plosive sounds are always followed by one of the Japanese five vowels. Ide et al[3] proposes a discrimination of Japanese voiceless plosives introducing time-spectrum pattern.

The authors[4,5] have proposed a discrimination method for CV syllables uttered isolatedly with plosive sounds as C followed by one of the Japanese vowels as V employing both instantaneous and dynamic properties of acoustic parameters at the burst and during the transition parts, respectively, where LPC cepstral coefficients and short time power are used as the acoustic parameters and their regression lines are introduced to represent their dynamics. All these experiments were conducted under condition that the given speech sample contains one of the plosive sounds.

This paper proposes an approach to detect plosive sounds in words. The proposed method employs LPC cepstral distance as the primary parameter to find out possible candidates for plosives. Also used as the auxiliary parameters for eliminating other phonemes from the candidate list are short time power, the pole positions obtained by low-order analyses, and a measure for representing relative disposition in the discrimination space.

This paper mainly discusses the process to eliminate excess errors of preliminary detection procedure. Phoneme identification or mutual discrimination among the detected plosive sounds is carried out by a following-vowel dependent discrimination algorithm formerly developed by the authors.

DETECTION OF PLOSIVE SOUNDS

Fig.1 shows the flow chart of the process from speech input to phoneme identification. The whole process is divided into two parts: one is detection of plosive sounds, and the other is mutual discrimination among them. In the detection part, possible candidates are first

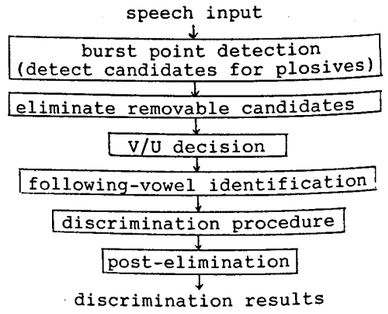


Fig.1 Flow chart of the process.

searched based on cepstral distance between an odd pair of frames with their starting points coincided. Then removable candidates are discarded from the candidate list in the elimination procedure. The rest of the process is related to discrimination among plosives. Followings are the detail of the processing blocks.

Burst Point Detection

Detection of plosive sounds can be realized by observing temporal changes of acoustic parameters. Temporal changes of acoustic parameters have been usually evaluated by comparing the difference between the analysis results of successive frames of the same frame length. However, the analysis scheme has problems of binary ambiguity and poor sensitivity because of its double-sided shifting gaps. In order to clear the these difficulties, an odd pair of frames with the starting points coincided as depicted in Fig.2 is introduced for detection of burst points or temporal change of spectral parameters. LPC cepstrum coefficient is employed here for the spectral parameter, and the distance d(t) between the analysis results of the odd pair of frames in the LPC cepstrum space is called cepstral distance where t denotes the center of the frame gap Lb.

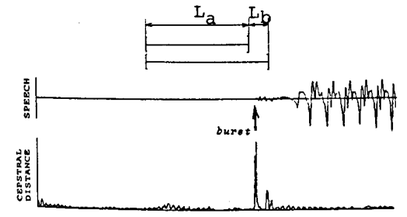


Fig.2 An odd pair of frames for detection of burst points.
Frame length: 20 ms for the longer frame
17 ms for the shorter frame
Shifting interval: 1 ms

Elimination of Removable Candidates

It seems impossible to detect plosive sounds only by the cepstral distance even though it is a good cue for burst point detection. Since the cepstral distance shows sharp peaks not only at burst points of plosive sounds but also at other occasions like beginning of fricatives, nasals and vowels. Auxiliary parameters introduced are short time power, pole position obtained by low-order LP analyses and a relative disposition of input sample in the discrimination space. The additional conditions to keep up as acceptable candidates are as follows:

a)Short time power P(t) grows up more than 40% of its local maximum P_m at the point t_p where the cepstral distance d(t) shows its local peak d_p.

$$P(t_p) / P_m > 0.4 \qquad (1)$$

where $d(t_p) = \max d(t)$

This is to eliminate the spectral transition part which shows a phenomenal local peak on the cepstral distance although it is not a plosive sound.

b)The average build-up rate r of the short time power P(t) is greater than a threshold r_o if t_p is at word-initial.

$$r = \frac{P_m - P_p}{t_m - t_p} > r_o \qquad (2)$$

This is to eliminate some phonemes which show slow power build-up in word-initial position like vowels.

c)LP analysis of order 2 yields no stable pole beyond 4kHz extending over successive 6 frames out of 28 frames shifted every 5ms around t_p.

This is to eliminate /s/ from the candidate list.

d)In case of voiced sounds, LP analysis of order 4 gives stable pole in the frequency range 80-300 Hz with narrower band-width than 200 Hz.

This is to eliminate /r/ from the candidate list.

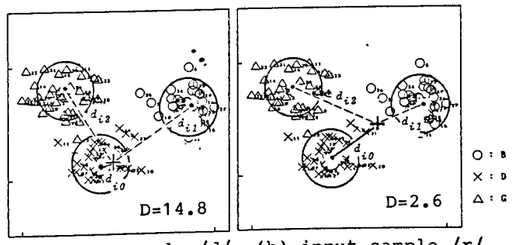
e)In word initial position, the following measure D, representing inclination to a particular plosive, is greater than a certain threshold D_o.

$$D = \sum_{i \neq i_o}^{N} \frac{d_i}{d_{i_o}} > D_o \qquad (3)$$

where di : distance from input sample to class i.

$d_{i_o} = \min_i d_i$

This is to eliminate phonemes of non-plosive disposition from the candidate list for plosive sounds. Fig.3 shows examples of relative disposition of the input samples /d/ and /r/ in the Fisher space with corresponding D values.



(a) input sample /d/ (b) input sample /r/
Fig.3 Relative disposition of /d/ and /r/ in the Fisher space.

This condition is applied after the discrimination procedure as post-elimination only to candidates in word-initial position.

The cepstral distance measure d(t) does not show any remarkable peak for sound /g/ in words since the temporal change of spectrum envelope is rather slow for the phoneme though it belongs to the plosive group. However, as /tʃi/ and /tsu/ yield evident peaks on the cepstral distance like plosives, the present report includes them in the set of phonemes to be detected and identified as /ti/ and /tu/, respectively.

DISCRIMINATION OF PLOSIVE SOUNDS

Fig.4 shows the discrimination process for the detected sounds.

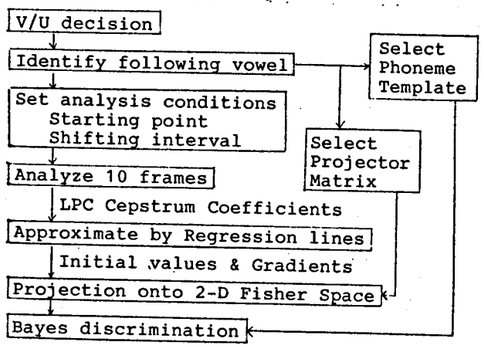


Fig.4 Discrimination among plosive sounds.

Voiced/Unvoiced Decision

V/U decision is made based on existence of buzz bar before the burst and the autocorrelation pattern of the prediction residual signal.

Identification of Following Vowels

Plosive sounds in Japanese are always followed by one of five vowels and the phoneme templates to discriminate plosives are prepared for each following vowel in the present system. So, identification of the following vowel should be performed before discrimination of the leading plosives. For identification of the following vowel two-stage decision by majority is employed. The first decision by majority is made among the 5-nearest neighbors in the discrimination space or a two-dimensional Fisher space derived from the LPC cepstrum coefficients, and the second decision by majority is made on the five decision results obtained from successive five frames shifted by 1 ms each, with the center frame at 70 ms after the burst.

Discrimination of Plosive Sounds

Both dynamic and instantaneous properties of acoustic parameters are employed for discrimination among plosives. Gradients of the regression lines representing temporal change of short time power and LPC cepstrum coefficients during the transition period from the burst to the following vowel are employed as the dynamic parameters. The number of frames to be fitted by regression lines is fixed to be 10. The first frame is located at the position fr1 starting at Td ms after the burst and the succeeding frames fr2, fr3, shifting interval Ts as depicted in Fig.5. The delay time Td representing the relative position of the first frame from the burst point and the shifting interval Ts are set optimal for each following vowel to give the best discrimination score for isolated CV utterances by preliminary experiments. The interpolated values on these regression lines at the specified positions are used as the instantaneous parameters.

The acoustic parameters here are C0, the frame power and C1, C2, ..., C12, the LPC cepstrum coefficients of order 12, and the parameters for discrimination among plosives are their gradients and the interpolated values for the first frame position. So the total number of parameters for discrimination is (1+12)x2=26.

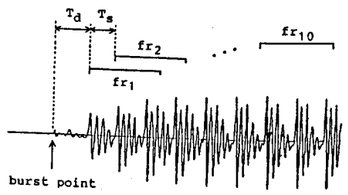


Fig.5 Set up of succeeding analysis frames.

Discrimination is performed in the two-dimensional Fisher space obtained from the 26-dimensional parameter space described above after V/U decision and following vowel identification. The phoneme templates are prepared for each following vowel and for voiced and unvoiced case respectively, that means there are 10 sets of phoneme templates containing three standard phonemes each. The resultant vector y in the Fisher space is obtained from the original parameter vector x as follows by the projector matrix W that maximizes the Fisher ratio J(W).

$$y = W^t x \tag{4}$$

$$J(W) = \frac{|W^t S_b W|}{|W^t S_w W|} \longrightarrow \max \tag{5}$$

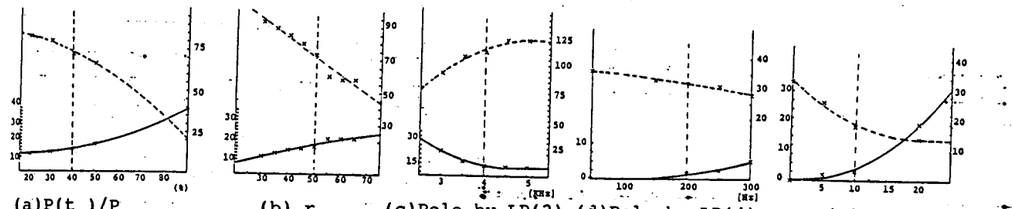


Fig.6 Performance of the parameters to eliminate removable candidates from the list.
solid lines : missing error dotted lines : excess error
vertical broken lines : threshold value

where t denotes transpose
S_w: within-class covariance matrix
S_b: between-class covariance matrix

Discrimination here assumes the normal distribution of each phoneme in the Fisher space and decision is made according to Bayes rule.

EXPERIMENTAL RESULTS

Speech Data

Phoneme templates in the Fisher space is obtained from isolated CV syllables uttered by 38 male adults. Test samples are 30 Japanese city names uttered by other 5 male adults. The number of phonemes to be detected and identified is 160 out of 535. Frequency of occurrence for each phoneme is not well balanced because of special feature of city names.

Results

Performance of the proposed method on above-mentioned test samples is shown in Table 1. It shows the statistics of missing errors and excess errors by adding conditions one by one on the test data.

Table-1 Detection and discrimination results of plosives in words. The total number of plosives to be detected = 160.
(0) LPC cepstrum distance $d(t) > d_o$
(a) Normalized power at the burst $^{o}P(t_p)/P_m > 0.4$
(b) Power build-up rate $r > r_o$ in word-initial
(c) Pole position in high frequency range by LP(2)
(d) Pole position in low frequency range by LP(4)
(e) Relative disposition $D > D_o$ in word-initial

Condition	Dis/Det/Tot	Missing	Excess
(0)	120/153/160	7	121
+(a)	116/147/160	13	71
+(b)	111/140/160	20	41
+(c)	111/140/160	20	36
+(d)	111/140/160	20	33
+(e)	109/137/160	23	18

Table-1 shows that the final detection rate is 86% (137 out of 160) with 15% excess-detection error, and 85%(109 out of 137) of the detected plosives are correctly discriminated.

Discussions about elimination conditions

Performance of the five parameters introduced in each elimination condition is shown in Fig.6 with a vertical broken line in each diagram representing the threshold value set in the present system.

CONCLUSION

Automatic detection and discrimination of plosive sounds are presented. LPC cepstral distance between an odd pair of frames is introduced as the primary cue for detection of plosives. Some other additional conditions are discussed to eliminate excess candidates for plosives.

Acknowledgment

This work is due to devotional help by Messrs.Mitsuhiro Tsunoda and Yukihiro Okada of Kansai University.

References

[1]S.Kitazawa et al.:J. Acoust. Soc. Jpn., 40, 5, 332-339(1984).
[2]M.Tominaga et al.: Trans. Committee on Speech Res., Ac.Soc. Jpn., S81-72(1982).
[3]K.Ide et al.: JAS Jpn, 39, 5, 321-329 (1983).
[4]Y.Yamashita et al. : Trans. IECE, Jpn., J69-A, 2, 282-290(1986).
[5]Y.Yamashita et al. : Trans. IECE Jpn., J70-A, 1, 132-134(1987).
[6]M.Tsunoda et al. : Autumn Meeting of Acoust. Soc. Jpn., 1-3-6(1986).
[7]Y.Takahashi et al. : Autumn Meeting of Acoust. Soc. Jpn., 2-1-5(1984).
[8]R.O.Duda and P.E.Hart:"Pattern classification and scene analysis", John Wiley, New York, pp.114-121(1973).