

CONTINUOUS SPEECH UNDERSTANDING BY KEYWORD EXTRACTION IN A VOICE MAIL SYSTEM

Y. Ariki*, H. Ohkawa^o. and T.Sakai*

ABSTRACT

In this paper, a developing voice-mail system is described which extracts keywords from continuously spoken mail. In this keyword extraction, a bottom-up approach to hypothesize the keywords and a top-down approach to verify them are integrated. First, a phoneme sequence is recognized in bottom-up mode, then keywords are hypothesized on the phoneme sequence. Finally, keyword candidates are verified by recognizing the consonants in a top-down manner.

INTRODUCTION

Recently, a voice mail system has been coming into wide use. This voice mail system accepts speech signals from telephone lines, then stores and exchanges them in a PBX. It greatly differs from a usual electronic mail system which deals with character codes, in retrieving or summarizing mail. In order to give such high capabilities to the voice mail system, speech recognition for voice mail is required. In this paper, a continuous speech recognition system is described for retrieving the contents of voice mail. In general, voice mail has various contents and the syntax is sometimes incomplete. It is, therefore, difficult to impose syntax constraints on voice mail. In order to accept various ways of speaking and to grasp the content of voice mail quickly, keyword extraction is carried out without using the syntax information.

APPROACH TO KEYWORD EXTRACTION

Type of Keyword Extraction

Keyword extraction methods may be classified into three groups. The first one extracts keywords by shifting the template frame by frame and matching it on a sequence of acoustic parameters of the continuous speech. Any highly correlated word and its corresponding location in the continuous speech are regarded as an output(ref 1). A continuous DP-matching is available to reduce the extraction time(ref 2). This method is highly accurate for a specified person, but costs much more processing time when the number of keywords increases. The second extracts keywords by template matching on a sequence of labels such as phonemes or pseudo-phonemes recognized in the continuous speech. Continuous DP-matching is also available. This method is suitable for speaker-independent and large number of keywords extraction. It also enables high speed extraction because the time sections with most plausible labels to the template are predicted and matching is performed within these predicted time sections. It is, however, inaccurate in extraction. The third extracts the candidates of the keywords on a sequence of labels, then verifies them on a sequence of the acoustic parameters(ref 3). This method locates between the first and the second. In this study, we employ the third method with the expectation of high speed and high accuracy extraction.

Integration of Top down and Bottom up Approach

The third method described above consists of three processes. The first is a bottom up process to recognize the label sequence from the acoustic parameters of the continuous speech. The second is a word hypothesis on the label sequence. The last is a top down process to verify the hypothesized words on the acoustic parameters. In the bottom up process, the recognition accuracy of the label sequence must be increased to reduce the number of the hypothesized keywords. On the other hand, in the top down process, the verification accuracy of the hypothesized keywords must be increased using the knowledge of the word. Fig.1 shows a block diagram of a developing keyword extraction system. The bottom up process and the top down

* Dept. of Information Science, Faculty of Engineering, Kyoto University, Kyoto, 606, Japan.

^o Computer Works, Mitsubishi Electric Corp. Kamakura-city, Kanagawa, 247, Japan.

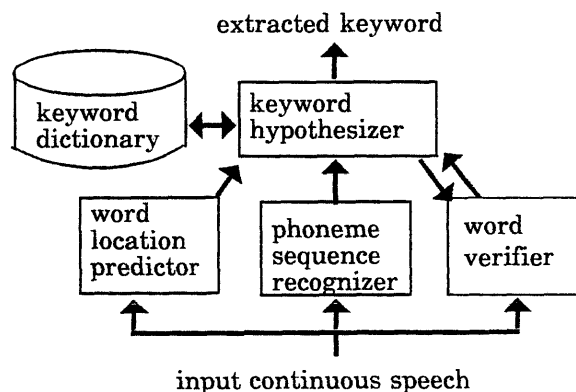


Fig.1 Block diagram of a keyword extraction system.

Table1 TDC analysis condition.

12 bits, 10kHz sampling
frame length: 25.6 ms
frame period: 6.4 ms
block length: 8 frames
block period: 4 frames
parameter: TDC coefficient
$\{C_{qp}, 0 \leq q \leq 14, 0 \leq p \leq 4,$
$(q,p) \neq (0,0)\}$

process are integrated via the keyword hypothesizer. In this system, phoneme recognition is performed to produce the label sequence.

Utilization of Acoustic Information to Predict the Word Location

A word location predictor in Fig.1 predicts the plausible location of the keyword using the acoustic information which the phoneme sequence recognizer has not utilized. The word hypothesizer extracts the keyword candidates within the predicted location on the phoneme sequence produced by the phoneme sequence recognizer. Therefore, the efficiency of the keyword extraction is increased by the predictor. At present, the power information is used as the acoustic information to predict the plausible location, because the power usually concentrates on the keyword to help the receiver to understand the meaning.

PHONEME SEQUENCE RECOGNIZER (BOTTOM UP PROCESS)

Two Dimensional Cepstrum (TDC)

The cepstrum parameters extracted from the short time period (frame) represent the static frequency structure. In phoneme recognition of the continuous speech, a longer time period is desired to extract the dynamic features as well as the static features. We employ the Two Dimensional Cepstrum (TDC) analysis for the phoneme recognition(ref 4). In the TDC analysis, several consecutive frames of the log spectrum are grouped as a block, and two dimensional FFT is applied to this. As a result, the two dimensional cepstrum coefficients are obtained for the block. The TDC coefficients have two axes p and q. The axis q corresponds to the quefrequency and has the time dimension. The axis p corresponds to the time frequency and has the frequency dimension. In the TDC coefficients, the static and dynamic features are represented separately as well as global and fine frequency structure. The continuous speech is at first represented as a time sequence of blocks and then converted to a sequence of TDC coefficients by TDC analysis. Table 1 shows the analysis condition.

Phoneme Recognition on the TDC

Fig.2 shows the phoneme recognition process. One phoneme label is assigned to one block. The most simple way of assigning phoneme label is realized by a linear discriminant function in multi class. It is, however, not so accurate due to the violation of normal distribution hypothesis of multi class. We employ multi-layer discrimination instead of one layer discrimination. At first, physical parameters of the TDC coefficients are statistically grouped into 256 classes (VQ classes) by a vector quantization technique(ref 5). These VQ classes may be regarded as the classes clearly separated in a perceptual space. Then, the VQ class is hierarchically mapped to the phoneme class. This multi-layer discrimination achieves the optimal mapping between the

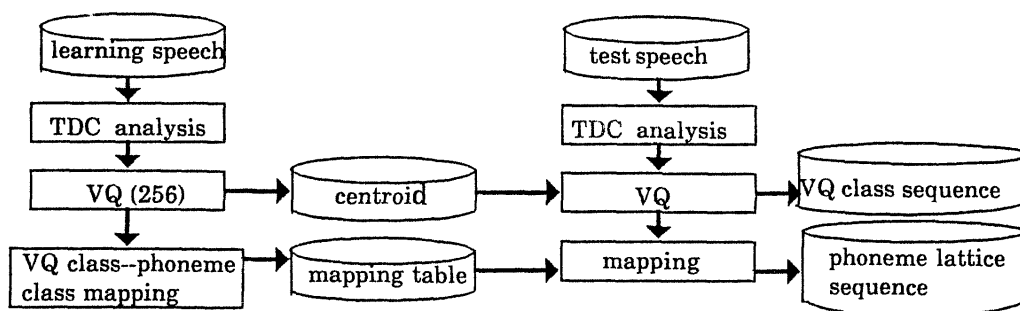


Fig.2 Recognition process of a phoneme sequence.

VQ class and the phoneme class. In Fig.2, 800 kinds of VCV learning speech data spoken by one male is analyzed by the TDC analysis and quantized into the 256 VQ classes. The same speech data is segmented into four parts $\{V_1, V_1C, CV_2, V_2\}$ and one of 32 phoneme classes shown in Table 2 is assigned to each of the four parts. In the mapping between the VQ class and the phoneme class, plural VQ classes may correspond to one phoneme class because the number of VQ classes (256) is greater than that of phoneme classes (32). In addition, one VQ class is allowed to be mapped to the plural phoneme classes when the VQ class is ambiguous in phoneme category.

Table2 List of the phoneme class.

a, i, u, e, o, N	6
p, t, k, s, h, b, d, g, z, r, w, y, m, n	14
ky, gy, sy, zy, cy, ny, ry, hy, by, py, my, silence	12

The test speech data is converted to a time sequence of the VQ class by the TDC analysis and the VQ. Then, by the mapping from the VQ class to the phoneme class, phoneme lattice sequence is produced with the score computed according to the distance to the centroids.

KEYWORD HYPOTHESIZER

On the recognized phoneme lattice sequence of the continuous speech, the keywords and their locations are hypothesized in the following manner:

- (1) **Word location prediction:**--Most plausible time sections for the keywords are predicted by the word location predictor in Fig.1. At first, time sections where the speech power is greater than a certain threshold are extracted. The end point of the time section corresponds to the end of syllables. Then, the time duration of the respective keyword is computed according to the number of syllables contained in the keyword. Finally, the consecutive time sections are connected until it exceed the time duration of the keyword. The connected time sections result in one most plausible time section for the keyword.
- (2) **Vowel extraction:**--On the phoneme lattice sequence, within the most plausible time section for the keyword, the block with the highest vowel score and its neighboring blocks with the same vowel labels are merged into the vowel sections.
- (3) **Word matching:**--The continuous DP-matching of the keyword is carried out on the vowel section sequence within the most plausible time section for the keyword. The keywords in the dictionary are represented in terms of the sequence of the phoneme class shown in Table2. The matching distance of vowels between the dictionary and input speech is hamming distance: 1 for accordance, 0 for inaccordance. The keyword and its location are determined if the total distance is greater than a certain threshold.
- (4) **Word scoring:**--If each consonant in the matched keyword is found on the phoneme lattice sequence at the consonant section, the consonant score is summed up. The word score is computed by summing up the vowel and consonant score and normalizing it by the number of syllables. The location and the word with score greater than a certain threshold are handed down as the hypothesized word to the verifier, the top down process.

WORD VERIFIER (TOP-DOWN PROCESS)

Each phoneme in the hypothesized keyword is verified by confirming that it locates at the hypothesized time location. For the top down verification, two class discrimination is superior in

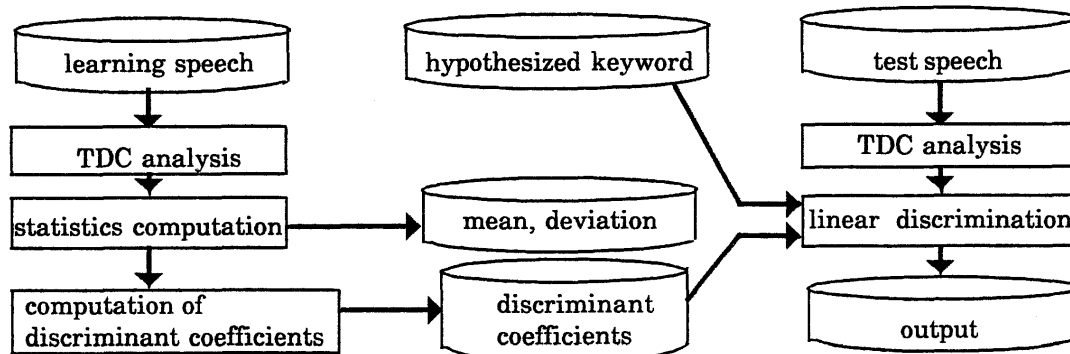


Fig.3 Phoneme verification process

terms of efficiency and accuracy. Here, we used the two class discrimination function: the phoneme and a group of the other phonemes. Fig.3 shows the process of phoneme verification. The coefficients of the two class discriminant function are computed from the learning speech. Using these coefficients, two-class linear discrimination of each phoneme in the hypothesized keyword is carried out on the test speech. The word is verified as the correct keyword if the likelihood of all the phonemes is greater than a certain threshold.

Table3 Keyword extraction rate.

	extraction rate	conformity rate
DP-matching	100.0	6.1
word hypothesis	100.0	32.7
top down	50.0	100.0

RESULT OF EXPERIMENT

An experiment of keyword extraction was carried out. Five words were selected as the keywords and were extracted from 20 sentences, in which each keyword appears four times. The time duration of the sentences is 15 second at average. Table3 shows the experimental result in terms of extraction rate and conformity rate of the keywords. The extraction rate is the ratio of the number of keywords correctly extracted to the number of keywords (20). This rate is a measure as to what degree the keywords are correctly extracted. On the other hand, the conformity rate is the ratio of the number of keywords correctly extracted to the total number of word extracted as keywords. This rate is the measure as to what degree the wrong words are extracted as keywords. Table3 shows both ratios after the continuous DP-matching on the vowel section sequence, after keyword hypothesis, and after keyword verification.

CONCLUSION

In this paper, a keyword extraction method for voice mail retrieval has been described. The main topics are: (1)Phoneme sequence recognition by applying the vector quantization technique to the TDC coefficients. (2)Keyword hypothesis by the continuous DP-matching on the vowel section sequence and by consonant verification. (3)Keyword verification by the two class linear discrimination of the phonemes in the hypothesized word. At present, the number of keywords is five (appearance is 20) for the 20 sentences. It will be extended to 100 words in the near future.

References:

1. R W Christiansen, IEEE Trans. on ASSP, vol.ASSP-25, no.5, (1977)
2. R Oka, ICASSP, (1986)
3. M F Medress, ICASSP, (1978)
4. Y Ariki, Proceedings of European Conf. on Speech Tech.,(1987)
5. Y Linde, IEEE Trans. on Comm., vol.com-28, no.1, (1980)