A Time Reduction Algorithm Using Vowel Classification for Large-Vocabulary Speech Recognition

Yong-Joo Jung and Chong Kwan Un

In this paper, a time reduction algorithm for large-vocabulary speech recognition has been studied. To reduce the recognition time, vowel was classified according to its formant information based on linear prediction analysis. Formant extraction was done in the assumed vowel regions. To segment vowel regions we used acoustic features such as energy, zero crossing rate and filter bank outputs. Computer simulation was done to test the time reduction algorithm with 1160 words. Simulation results show that the recognition rate is about 97% and the time for recognition can be reduced by 80% as compared to the case without the reduction algorithm.

I. Introduction

In large-vocabulary speech recognition, it is normally necessary to use a preprocessor to reduce the recognition time. The purpose of the preprocessor is to reduce the number of candidate words for the input word. In our speech recognition system, the vocabulary consists of 1160 words. Therefore, it takes excessive time to process all the words in the main part if a preprocessor is not used.

One method to implement the preprocessor is to use the vector quantization (VQ) approach (Shors & Burton (1983)). In this method, every word in the vocabulary has a VQ codebook and it is compared with the codebook of the input word. The words with large distance from the input word are excluded from the candidate words. Then the words that survive are the input of the main processor. However, this method requires much time and memory to construct VQ codewords. The other method is to use acoustic-

Language Research, Volume 27, Number 2, June 1991. 0254-4474 / 219-227

phonetic features of speech (Chen (1986)). The acoustic-phonetic features can be the characteristic of each phoneme. They can be represented by filter bank outputs, zero crossing rate, and so forth. This method does not require memory in proportion to the vocabulary size, but it is not easy to find the acoustic-phonetic features which uniquely characterize each phoneme.

Here we use the latter approach and implement the preprocessor which classifies the vowels in the input word. Fig. 1 shows the structure of our large-vocabulary speech recognition system. The preprocessor consists of the vowel-classification part and the word-classification part. The vowel of an input word is classified according to its formant values by the vowel-classification part. The classification of vowels in the input word gives a clue of determining the candidate words from the dictionary. Since the formant value is obtained in the assumed vowel regions, it is necessary to segment the input speech correctly.

The procedure of segmentation is discussed in the next section. The method to obtain formant values and the implementation of the word-classification part are considered in Section Ⅲ. Finally, simulation results are given in Section Ⅳ.

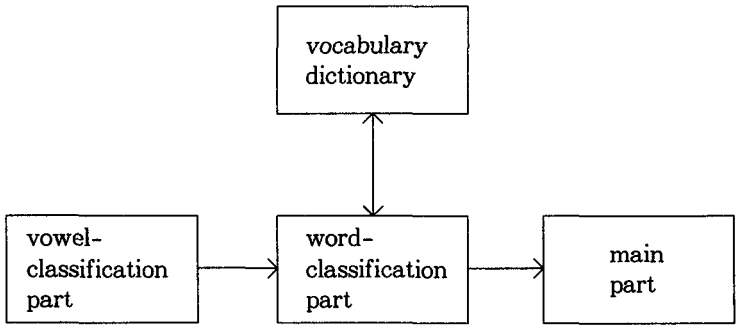

Fig. 1. Processing in a Large-Vocabulary Speech Recognition System.

II. Segmentation

To segment vowel regions, we first obtained zero crossing rate, energy and filter bank outputs (125~750 Hz, 1~3 KHz, 4~5 KHz). We call the output of 125~750 Hz filter bank as fbl, the output of 1~3 KHz filter

bank as fb2 and the output of 4~5 KHz filter bank as fb3. To use these feature values, we assumed five regions; high, low, edge, dip and fast change regions. High region means that the feature value is high relative to other parts of the input speech. We used two threshold values, T1 and T2 which are the starting threshold value and the ending threshold value of high region. We set T1 high enough to ensure the start of high region and T2 low enough to take account for transient perturbation effects in the true high region (Chen (1985)). That is,

$$T1 = C_1 (\max_{\mu} - \min_{\mu}) + \min_{\mu}$$
$$T2 = C_3 (\max_{local} - \min_{\mu}) + \min_{\mu}$$

where,

$$\max_{\mu} = \begin{cases} \max_{observed} & \text{if } \max_{observed} > \min_{global} \\ -r(\max_{global} - \min_{global}) & \\ \max_{global}, & \textit{otherwise} \end{cases}$$

$$\min_{\mu} = \begin{cases} \min_{observed} & \text{if } \min_{observed} < \min_{global} \\ +r(\max_{global} - \min_{global}) & \\ \min_{global}, & \textit{otherwise.} \end{cases}$$

As can be seen above, T1 and T2 are determined from the maximum ($\max_{observed}$) and minimum values ($\min_{observed}$) of input speech and the predetermined maximum and minimum values ($\max_{global}$, $\min_{global}$). $\max_{global}$ and $\min_{global}$ are determined from the average magnitude of normal speech. After T1 is found, the feature value is smoothed to account for temporal perturbations as follows.

$$S[i] = C_2 X[i] + (1-C_2) S[i-1],$$

where $X[i]$ is a raw feature value and $S[i]$ is a smoothed value.

T2 is also affected by the peak of the smoothed feature value ($\max_{local}$). Low region is determined by a method similar to the method used for the high region with only slight modification.

Dip region is where the feature value is relatively high and steady transients of feature values appear. To find the region, we obtained 3-point median, 7-point median filtered values. From these smoothed values, we find the point where the slope of values changes from negative to positive. On each side of the point, a local maximum value is found (at this point the

slope changes from positive to negative), and if the difference between this local maximum value and the value of the first point selected is larger than some threshold, the selected point is counted as a dip point. Accordingly, we obtain two kinds of dip points, 7-dip point and 3-dip point. This dip point gives as good clue for finding nasal sound.

Edge region is used to discriminate voiced and unvoiced sound. If two neighbor frames have larger difference in value than some threshold, they are regarded as the start of an edge region and edge region is regarded as a voiced region. This is reasonable because the energy value in a voiced region has large fluctuation between frames.

If the difference of the feature values between two frames which are 3 frames away is larger than a threshold value, they are assumed to be the beginning of a fast change region. The fast change region is used to find the boundaries of phonemes.

We next discuss how we can use the above five regions of each feature to segment vowel from the input speech. Fig. 2 shows the procedure to detect vowel regions from the input speech. In a vowel region, the fb1 energy is typically large and the total energy is also large. Therefore, one can roughly find vowel regions by selecting high regions of fb1 energy and the total energy from input speech.

The region with large fb1 energy is normally a vowel region, but the region with large total energy can sometimes be caused by consonant with much high frequency energy. Therefore, from the assumed vowel regions, we exclude the regions of large fb3 energy and high zero crossing rate.

Nasal sound has large fb1 and total energy. Therefore, it is often confused with vowels. But nasal sound has many dip points in fb1 and fb2 energy. By detecting dip points in fb1 and fb2 energy, we can find nasal sounds to exclude them from the assumed vowel regions.

Edge region is used to detect unvoiced sound which has large fb1 or total energy. Fast change region is used to determine phoneme boundaries.

In the final step, the length of vowel regions is checked. If it is too long, it is reexamined by the above procedure with different threshold values or it is assumed as a cascade of two vowels. If the inter-distance of vowel regions is too long, another vowel is assumed between two vowel regions. The final result is a segmentation of vowel regions from input speech. If the correct vowel region is not obtained by the above procedure, the existence informa-

tion of vowel can also be a clue in the input word classification scheme.

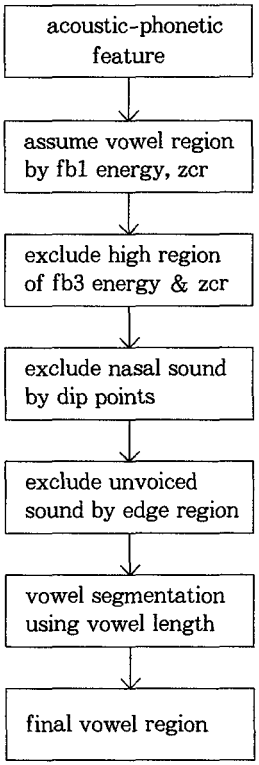

Fig. 2. Procedure of Determining Vowel Regions.

III. Formant Extraction

Once a vowel region is obtained as discussed in Section II, we can extract formant values from the vowel region. The procedure to extract formant values is as follows. First, linear prediction coding (LPC) coefficients are obtained for every frame of input speech. From these coefficients, we can obtain spectral magnitudes and peaks. These peaks can be thought of as formant values of the vowel region (Markel & Gray (1976)).

Among these raw formant values, there are often incorrect values. This is because LPC analysis is based on an all-pole model. Especially, in a nasalized vowel sound, it is difficult to find exact formant values because

nasal sound is affected by zeros in the vocal tract transfer function. Some formant values are often cancelled and extra formant values appear. The reason of the above difficulty is that poles inside the unit circle is not explicitly found by spectral calculation on the unit circle. For this case, we found the spectrum on a circle with radius less than 1. The radius of circle was reduced from 1 to 0.88 in trying to find reasonable formant values. If we reduce the radius of circle too much, the peaks caused by poles near the unit circle in the z-transform domain becomes less dominant. Therefore, we did not reduce the radius less than 0.88.

To obtain reliable formant values, each formant value in one frame is compared with the formant values of the preceding frame. If the extracted formant values in the present frame are not much different from those of the preceding frame and are reasonable from the viewpoint of average formant values, they are chosen as the formant values of the present frame. If not, the formant values of the present frame are determined based on the formant values of the preceding frame. From the above procedure, the first three formant values (F1, F2, F3) are obtained in each frame.

In our formant analysis, the length of each frame is 18 ms and the window was slided down by 9 ms each time. Since the length of vowel regions range about 30~150 ms, there are 3~16 frames in each vowel region. From the many formant values, we must extract the representative value of each vowel, because each vowel has formants that characterize it. In this case, a procedure, such as averaging or excluding extraordinary values, is required.

To classify vowels, we used only F1 and F2. With F1 and F2 values, vowel is classified in three categories. The first category is the one that has high F1, typically more than 520 Hz. The second category are vowels that have low F1 and relatively low F2. In this case F1 is less than 390 Hz and F2 is less than 1600 Hz. The last category are vowels with F2 larger than 1600Hz. Although we can classify vowels as above, formant values differ very much according to different speakers and are influenced by the position of vowels in the input speech. The same vowel may have different formant values depending on whether the position is in the front part or the latter part of a word or it is near a nasal sound. Since the difference among speakers is large, we experimented with the input speech of only one speaker. By taking into consideration the above effects, we made a dictionary for 1160 words to use them in the time reduction part of a recognition system. In the dictionary, each word is labelled by the class of vowels in the word. That is, if a word has three vowels and the category of vowels is A, B, C,

respectively and sequentially, then the dictionary content for the word is 'ABC'.

We can briefly explain the time reduction part of our system as follows. If a word is the input of our system, the system segments it, extracts formant values from the assumed vowel regions, and classifies each vowel. The classified vowel sequence is the characteristic of the input word. If the classified vowel sequence is A, B, C, we search for the word in the dictionary with label 'ABC'. The other words in the dictionary which do not have this label are excluded from the candidate words for the given input word. Only the candidate words are passed to the main part of the system. If the desired word(input word) is not contained in the passed candidate words, it is counted as a recognition error because it cannot be recovered in the main part. The average ratio of excluded words in the dictionary for each input word to the vocabulary size (i.e., 1160 in our system) is the time reduction rate. For example, if we pass 200 candidate words for each input word in average, the time reduction rate is 960/1160.

The next section shows the simulation result.

Ⅳ. Simulation Result

We experimented the above time reduction algorithm for one male speaker. The database consists of five groups; S1, S2, A1, A2, and A3. Each of S1 and S2 consists of 80 sentences which were constructed based on the 1160 words. One sentence consists of 5~7 words. Each of A1, A2, and A3 consists of the entire 1160 words. The simulation result is shown in Table 1.

Since the time reduction part is a preprocessor of the recognition system, if an error occurs, it cannot be recovered. There are largely two causes of errors. One is due to segmentation errors. If a vowel region is not exactly segmented, the later formant extraction process becomes meaningless. The other is due to the failure of extracting correct formant values. Also the construction of a dictionary is important, since it can compromise sound effects in real speech. In our experiment, we found that nasaling of vowels results in errors most frequently. It makes segmentation and formant extraction very difficult. Therefore, it appears that unless the preprocessor performs accurately, the time reduction part is not effective when the required recognition rate is very high. But we can make it useful to some extent by utilizing more features.

Table 1. Result of the Time Reduction Test.

Data Base	Recognition Rate	Time Reduction Ratio
S1	96%	85%
S2	98%	87%
A1	97%	80%
A2	97%	79%
A3	96%	79%

V. Conclusion

We proposed the use of a preprocessor for large-vocabulary speech recognition system. The preprocessor utilizes acoustic-phonetic features to segment vowel regions of the input speech. And formant values in the assumed vowels are obtained to specify each input word. This process reduces the candidate words for the input word. It is important that the required input word should be included in the candidate words with high accuracy. If an error occurs in the preprocessor, it cannot be recovered in the later part of the recognition system. The simulation results show that the recognition rate is about 97% and the time for recognition can be reduced by 80% as compared with the system without the reduction algorithm.

References

Andre-Obrecht, R. (1986) 'Automatic Segmentation of Continuous Speech Signals,' *Proc. IEEE Int. Conf. ASSP.*

Chen, F. R. (1985) 'Acoustic-Phonetic Constraints in Continuous Speech Recognition: A Case Study Using the Digit Vocabulary,' Ph. D. dissertation, Massachusetts Institute of Technology.

Chen, F. R. (1986) 'Lexical Access and Verification in a Broad Phonetic Approach to Continuous Digit Recognition,' *Proc. IEEE Int. Conf. ASSP.*

Cole, R. A. and L. Hou (1988) 'Segmentation and Broad Classification of Continuous Speech,' *Proc. IEEE Int. Conf. ASSP*, pp. 435-456.

MacCandless, S. S. (1974) 'An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra,' *IEEE Trans. ASSP*, Vol. 22, pp. 135-141.

MacCandless, S. S. (1976) 'Modifications to Formant Tracking Algorithm of April 1974,' *IEEE Trans. ASSP*, Vol. 24, pp. 192-193.

Markel, J. D. (1971) 'Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation,' *IEEE Trans. Audio and Elec.*, Vol. 20, pp. 129-137.

Markel, J. D. and A. H. Gray, Jr. (1976) *Linear Prediction of Speech*, New York, Springer-Verlag.

Pan, K. C. et al. (1985) 'An Efficient Vector Quantion Preprocessor for Speaker Independent Isolated Word Recognition,' *Proc. IEEE Int. Conf. ASSP*, pp. 874-877.

Rabomer, L. R. and R. W. Shafer (1978) *Digital Processing of Speech Signals*, Englewood Cliffs, N. J., Prentice-Hall.

Shors, J. E. and D. K. Burton (1983) 'Discrete Utterance Speech Recognition without Time Alligment,' *IEEE Trans. Information Theory*, Vol. 29, No. 4, pp. 473-491.

Center for Speech Information Research
Communications Research Laboratory
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
P. O. Box 150, Chongyangni, Seoul
Korea