한국통신학회 1987년도 추계학술발표회 논문집 (1987. 11. 14)

변형된 거리척도에 의한 음향학적으로 유사한 단어들 사이의 변별력 개선

°김 형 순, 온 종 관
한국과학기술원 전기 및 전자공학과

On Improving Discriminability among Acoustically Similar Words by Modified Distance Metric

Hyung Soon Kim and Chong Kwan Un
Department of Electrical Engineering, KAIST

ABSTRACT

In a template-matching-based speech recognition system, excessive weight given to perceptually unimportant spectral variations is undesirable for discriminating among acoustically similar words. By introducing a simple threshold-type nonlinearity applied to the distance metric, the word recognition performance can be improved for a vocabulary with similar sounding words, without modifying the system structure.

1. INTRODUCTION

Template matching is known to be one of the most popular and successful approaches to speech recognition, and most commercial speech recognition systems are based on this approach. In template matching, each word is represented by a sequence of spectral patterns as a function of time. Recognition is done by comparing the unknown input against each of the stored word templates using a predefined distance metric. To account for some variability in speaking rate, a time normalization procedure known as dynamic time warping is usually used.

One of the major drawbacks of template matching is that it gives equal attention to all time frames[1]. Since total word matching scores are computed by adding successive frame-wise local distances, excessive weight is given to perceptually unimportant frame-to-frame variations in long-duration stationary vowels. Consequently, while a template matching system yields high accuracy for vocabularies with acoustically distinct words, its performance becomes degraded for vocabularies with similar sounding words. As solutions to this problem, the two-pass approach[1] and the discriminative network approach[2] have been proposed. However, they are very complex, and furthermore the change of vocabularies is not easy in those approaches because of their vocabulary-dependent structure.

In this paper, we propose a simple technique in which only distance metric is modified without affecting the whole system structure. By applying a threshold-type nonlinearity to the distance metric used, the effect of perceptually unimportant variations can be reduced, thereby increasing the recognition accuracy, especially for the vocabularies with acoustically similar words.

2. DESCRIPTION OF THE PROPOSED METHOD

Typical curves of local distance over the optimal warping path (based on dynamic time warping) versus time frame are shown in Fig.1. In this figure, solid and dotted curves indicate correct word matching and incorrect but acoustically similar word matching cases, respectively. For the incorrect word matching case, local distance values in the phonetically different region(*region A* in Fig.1) are greater than those in the phonetically identical region(*region B* in Fig.1). The total matching score is represented by the area under each curve. One can note in this figure that, since local distances due to perceptually unimportant spectral variations of the correct word matching case are relatively larger than those of the incorrect word matching case (*region B* in

Fig.1), and the size of *region A* is much smaller than that of *region B*, the score of incorrect word matching may be lower than that of correct word matching, thus causing wrong classification.

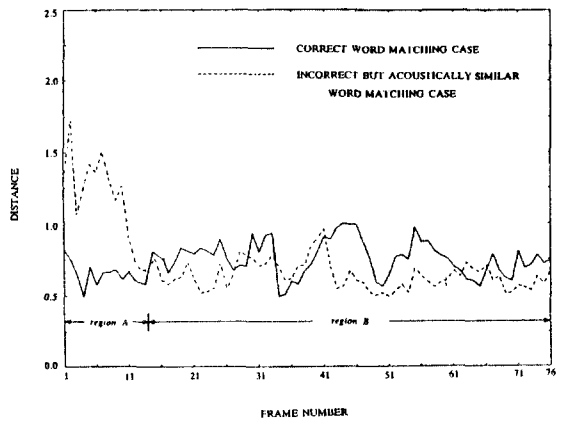

그림 1 Dynamic time warping에 의한 최적 warping 경로상에서의 local distance의 전형적인 곡선.

Fig. 1 Typical curves of local distance over the optimal warping path (based on dynamic time warping) versus time frame.

If the local distance metric is modified so that the effect of perceptually unimportant variations is reduced, the recognition accuracy can be improved. For this purpose, we apply a threshold-type nonlinearity to the local distance metric. The modified distance metric, $\hat{d}(\cdot,\cdot)$, is obtained from the original distance metric, $d(\cdot,\cdot)$, as the following:

$$\hat{d}(\cdot,\cdot) = f[d(\cdot,\cdot)] \tag{1}$$

where

$$\begin{aligned} f[x] &= x - TH \quad , \text{ if } x > TH \\ &= 0 \quad\quad , \text{ otherwise} \end{aligned} \tag{2}$$

and TH is an appropriately chosen nonnegative constant. If the TH value is too small, the effect of irrelevant spectral changes is not reduced. On the other hand, if the TH value is too large, perceptually important changes are also neglected and word discrimination becomes impossible. We select the optimal TH value experimentally under the criterion of recognition accuracy.

It is important to note that the above type of nonlinearity is not the only possible choice. Our purpose here is mainly to demonstrate the possibility of improving the discriminablity among similar sounding words through some threshold-type nonlinear weighting on the distance metric. Another type of modified distance metric has been proposed by Haltsonen[4], where the original distance is raised to power p. If p is greater than one, larger distance values are weighted more heavily. Experimental results including a comparison of these two schemes follow.

3. EXPERIMENT AND DISCUSSION

An experiment has been performed for a vocabulary of 62 Korean geographical names, which contains considerable amount of acoustically similar words. Average number of syllables in the vocabulary is 2.4. Five repetitions were pronounced by eight male speakers. The first repetition of each speaker was chosen as a reference set in the speaker-dependent mode, and the remaining four were used for testing. Thus, a total of 1984 test tokens was obtained.

Spectral analysis of incoming speech was carried out every 10 ms by a bank of 17 critical-band-spaced filters, ranging from 130 to 4300 Hz. After logarithmic transformation, amplitude normalization by mean was performed to compensate for variations in speech level. Word boundaries were automatically detected based on the frame energy and zero crossing rate.

The word recognizer used in this experiment employed the dynamic time warping with type-*Id* local constraints of [3] and parallelogram-type global constraints. As a basic distance metric for spectral distance computation, the Euclidean distance metric was chosen, which resulted in slightly better performance than the absolute distance metric under our experiment condition.

Our simulation results are shown in Fig.2. As shown in this figure, the threshold-type nonlinearity applied to the distance metric decreases the error rate in a relatively wide

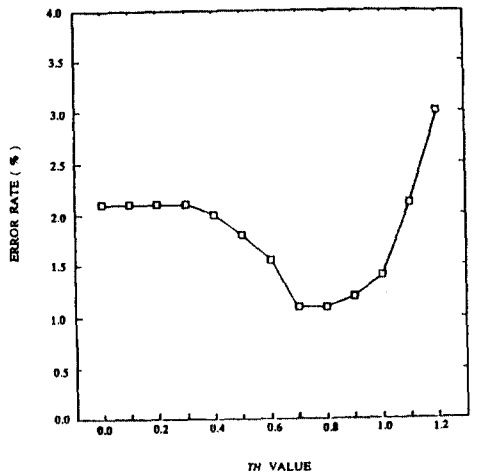

그림 2 *TH* 값의 변화에 따른 오인식율.

Fig. 2 Recognition error rates as a function of *TH* value.

range of *TH* values. For *TH* =0.7, in particular, the error rate is reduced to 1.1 percent, which is about one half of the original error rate(2.1 percent for *TH* =0.0). From the vocabulary of 62 words, most errors(90 percent) occurred in 15 highly confusable words. For these words, the error rate is reduced from 7.7 percent (for *TH* =0.0) to 3.8 percent (for *TH* =0.7), and for the remaining words, the error rate is also decreased slightly. This means that the proposed scheme yields improvement in discriminating among acoustically similar words without affecting discrimination between acoustically dissimilar words. Results for the distance metric raised to power p as studied in [4] have also been obtained. Even with the parameter p optimally chosen (p =2.8 in this experiment), the error rate was 1.5 percent for full 62 words and 5.6 percent for 15 highly confusable words, which is higher than that of the proposed scheme. This may be due to the fact that the threshold-type nonlinearity method is more effective in reducing the effect of irrelevant spectral variations.

Fig.3 shows the distributions of local distances before applying nonlinearity. In Fig.3, (a) and (b) are the distributions of local distances with correct word matching and incorrect word matching cases, respectively, and (c) is the local distance distributions of incorrect but acoustically the most similar word matching case. While (b) and (c) contain both phonetically identical and phonetically different frame matching, (a) can be regarded as the distribution of local distance with phonetically identical frame matching. Comparing Figs.2 and 3, it can be observed that the error rate is reduced when the *TH* values are in the range where local distances of phonetically identical frame matching are distributed, and that the error rate is increased abruptly when the *TH* value exceeds this range.

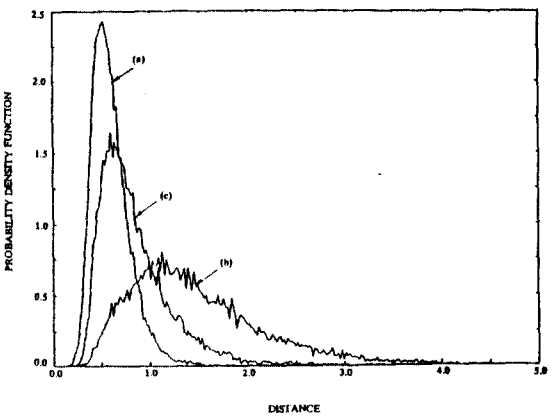

그림 3 최적 warping 경로상에서의 local distance 값의 분포. (a) 올바른 단어 matching의 경우. (b) 틀린 단어 matching의 경우. (c) 틀린 단어이나 음향학적으로 가장 유사한 단어 matching의 경우.

Fig. 3 Distributions of local distances over the optimal warping path.
(a) Correct word matching case. (b) Incorrect word matching case. (c) Incorrect but acoustically the most similar word matching case.

One can note that although the performance can be increased by this simple technique, more refinement should be done. The assumption behind this approach is that spectral variations below certain threshold are not information-bearing from the human perception viewpoints. However, since the spectral variability of speech data is not independent of the specific sound and word in the vocabulary[5], the use of frame-specific thresholds based on the training procedure can

yield better performance. Further research on other forms of nonlinearity and frame-specific thresholds is being done.

REFERENCES

[1] L. R. Rabiner and J. G. Wilpon, "A two-pass pattern-recognition approach to isolated word recognition," Bell Syst. Tech. J., vol.60, pp.739-766, May-June, 1981.

[2] R. K. Moore, M. J. Russell and M. J. Tomlinson, "The discriminative network ; a mechanism for focusing recognition in whole-word pattern matching," in Proc. 1983 IEEE Int. Conf. Acoust., Speech, Signal Processing, Apr. 1983. pp.1041-1044.

[3] C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-28, pp.622-635, Dec. 1980.

[4] S. Haltsonen, "Improved dynamic time warping methods for discrete utterance recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-33, pp.449-450, Apr. 1985.

[5] E. L. Bocchieri and G. R. Doddington, "Frame-specific statistical features for speaker independent speech recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-34, pp.755-764, Aug. 1986.