Robust DTW-based Recognition Algorithm for Hand-held Consumer Devices

Chanwoo Kim and Kwang-deok Seo
LG Electronics Inc., 60-39, Kasan-dong, Kumchon-gu, Seoul, 153-801, Republic of Korea

Abstract-- **This paper presents a new Dynamic Programming (DP)-based recognition algorithm that is quite suitable for menu-driven recognition applications with small vocabulary size (typically less than 50). In terms of computational complexity and parameter file size, the proposed algorithm shows improved results for command recognition system with small vocabulary size than the conventional Hidden Markov Model (HMM)-based recognizer. In addition, the proposed algorithm shows much higher recognition accuracy compared to the conventional Dynamic Time Warping (DTW)-based recognizer.**

I. Introduction

Dynamic programming and its modifications have been successfully adopted for speech recognizer. This type of recognizer is commonly employed in handheld consumer devices like cell phones in the form of Dynamic Time Warping (DTW). Specifically, DTW-based recognition engine has been widely embedded inside Qualcomm MSM (mobile station modem) chips [1]. However, an inherent problem found in DTW algorithm is that it is vulnerable to speaker-independent (SI) recognition case [2].

Although HMM-based recognition engine has just begun to be employed for handheld devices due to its advantages in large vocabulary size and continuous speech recognition, DTW method still has various applicable areas including menu-driven commanding and phone dialing due to its low computational complexity. Moreover, in the case where the number of reference patterns is small, the required DB size needed for DTW method can also be kept small.

In this paper, we propose a new method that circumvents several inherent shortcomings of the conventional DTW algorithm. For this purpose, DP algorithm is employed in two levels: lower level and upper level. In the lower level, the algorithm is quite similar to the one that is used in the conventional DTW method while in the upper level we try to match the chunks of speech by using DP in order to obtain improved time alignment, thereby resulting in much higher recognition rate. Through extensive simulations using 100 sentences, the proposed algorithm shows improved recognition rate compared to the conventional DTW algorithm.

II. Proposed Algorithm

The flow diagram of the proposed algorithm is illustrated in Fig. 1. Basically, this algorithm partitions input speech into chunks of speech and applies the DP in two-levels. Specifically, feature extraction and node distance computations are done first as shown in Fig. 1. Thereafter, speech is partitioned into voiced, unvoiced and silence parts and the beginning and ending points of each chunk are marked. We perform the partitioning on both the reference and test speech patterns. Then we compute the partitioned global distances in the lower-level DP. In the upper-level DP, we employ another DP to align the transition points and to compute the total global distance.

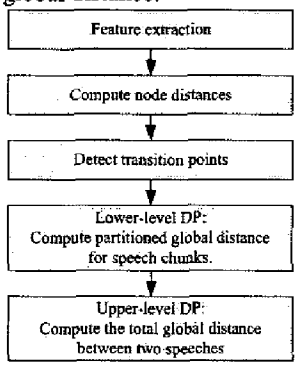


Fig. 1. Flow diagram of the proposed algorithm

A. DP Algorithm in Lower Level

After finding the transition points in the test and reference speeches based on speech/silence and voiced/unvoiced information, lower-level DP is performed. The following equation (1) is employed in this stage:

$$D_{acc}(i'_{test}, i'_{ref}) = \min_{i_{test}, i_{ref}} \{D_{acc}(i_{test}, i_{ref}) + d((i'_{test}, i'_{ref}), (i_{test}, i_{ref}))\} \quad (1)$$

where i'_{test} and i'_{ref} denote the current frame indices of test and reference patterns, respectively. And i_{test} and i_{ref} denote the previous frame indices of test and reference patterns, respectively. Note that (1) is used to compute the distance between partitioned chunks of speech.

B. DP Algorithm in Upper Level

After computing the partitioned global distances using (1), we compute the total global distance by the best-matched transition points. This matching process is performed by an additional DP. For this stage, we use another plane of grids where each grid denotes the beginning and ending points of each speech chunks. The following equation is employed for this stage:

$$D_t(I', J') = \min\{D_t(I, J) + D_{acc}((I, J), (I', J'))\} \quad (2)$$

where $D_{acc}((I,J),(I',J'))$ is the partitioned global distance between the test and reference chunks of speech. In (2), I and J are the frame indices of *I-th* transition point of the test pattern and *J-th* transition point of the reference pattern, respectively. And the prime notation in (2) means the specified indices are current ones whereas indices without it are previous ones. $D_t(I,J)$ means the sum of the partitioned global distances computed at this transition point along the

0-7803-8838-0/05/$20.00 ©2005 IEEE.

path obtained by DP.

An example of the proposed partitioning and alignment is shown in Fig. 2 where I_{num_trans} and J_{num_trans} denote the number of transition points in the test and reference speeches, respectively. The solid curse shows a path alignment within a specific partitioned speech using (1). The entire path comprised of both the solid and dotted curves are obtained using (2). Note that each grid does not designate a frame of speech but a transition point. After repeated experiments, we adopt (3) as the local continuity constraint in the upper-level DP. Strict constraint as used in the conventional DP often leads to degraded performance in this DP.

$$\begin{aligned} 1 \le |I - I'| \le 4 \\ 1 \le |J - J'| \le 4 \end{aligned} \tag{3}$$

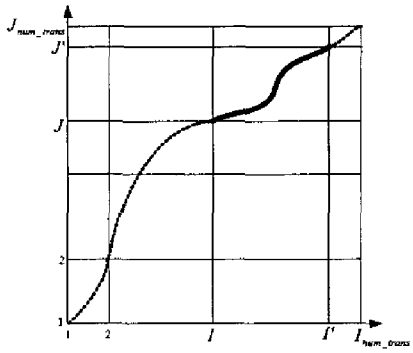
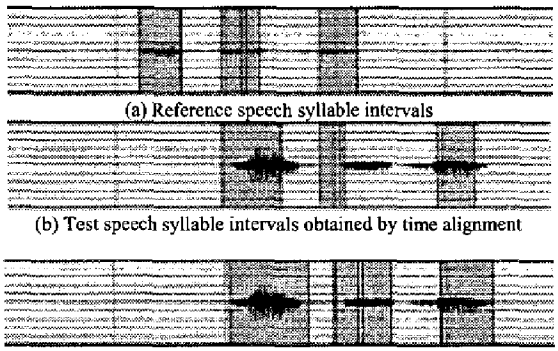


Fig. 2. DP in upper-level transition points

III. Experimental Results

To compare the proposed algorithm with the conventional DTW recognizer, the proposed algorithm is tested for 100 sentences spoken by 5 speakers. Recognition accuracies for inter- and intra- speaker cases are evaluated. In SI recognition case, the proposed algorithm shows a remarkable improvement as shown in Fig. 3. Fig. 3(a) shows the reference speech and the shaded regions in this figure are syllables marked by a human. Fig. 3(b) and Fig. 3(c) show the regions in the test speech that correspond to these shaded regions in the reference speech. We use the conventional DTW-based algorithm and the proposed algorithm in case of Fig. 3(b) and Fig. 3(c), respectively. In SD cases, the improvement in speech recognition accuracy is not noticeable, whereas the improvement reaches up to 5% for the SI cases. This is due to the fact that the recognition accuracy using the conventional DP algorithm is sufficiently high and consequently there is little headroom for further improvement. Considering the 5 speakers include a female, our system also shows robustness for inter-sex speech recognition cases. Table 1 shows the recognition rate when different types of features like LPCC, MFCC and PLP are used for 100 speech samples. In order to obtain various features in the evaluation phase, we used a feature extraction tool named HCopy that is a member of speech recognizer suite HTK[3]. For various features, the improvement ranges between 2 to 4%. From these results, we can conclude that the proposed algorithm is much superior to the conventional DTW algorithm.

For another comparison, we performed the same test using a simple triphone HMM-based recognizer and obtained the recognition rate of 92% for the same test set, which is almost the same as that of the proposed system. Although HMM-based algorithm still shows higher accuracy in SI cases, the accuracy difference is reduced significantly compared to the conventional DTW-based algorithm. Moreover, in SD cases, the proposed algorithms shows better results compared to HMM-based one. While HMM-based recognizer needs a parameter file size of 748 KB, the proposed algorithm requires only 100 KB in this experiment. And for 10 vocabulary cases, the computational amount needed by the proposed algorithm is less than half of that is needed by the triphone HMM-based recognizer. According to this fact, the proposed algorithm is quite suitable for hand-held consumer devices with limited memory and CPU resources.



(a) Reference speech syllable intervals

(b) Test speech syllable intervals obtained by time alignment

(c) Test speech syllable intervals obtained by time alignment

Fig. 3. Improved time alignment by the proposed algorithm

TABLE I.
RECOGNITION RATE COMPARISON FOR 100 SPEECH SAMPLES

Feature Type	Proposed Method	Conventional Method
LPCC	0.87	0.83
MFCC	0.89	0.86
PLP	0.90	0.88

IV. Conclusions

In this paper, we propose a robust DP-based speech recognition algorithm that is quite suitable for menu-driven recognition applications with small vocabulary size. The proposed algorithm shows improved robustness for SI cases compared to conventional DTW-based recognizers. The fact that it requires light computational complexity and small parameter file size makes it suitable for applications with small vocabulary size in hand-held consumer devices.

References

[1] Qualcomm, PureVoice™ VR Toolkit Version 2.1.0 User Guide and Script Language Description, San Diego, CA: Qualcomm, Mar. 2001.

[2] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice Hall, 1993.

[3] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book (for HTK Version 3.2). Cambridge, UK: Cambridge University Engineering Dept., 2002.