THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.26, NO.1E 2007. 3 pp. 21~26

A Modified Viterbi Algorithm for Word Boundary Detection Error Compensation

Hoon Chung*, Ikjoo Chung**
*ETRI, **Kangwon National University
(Received January 24 2007; Accepted March 19 2007)

Abstract

In this paper, we propose a modified Viterbi algorithm to compensate for endpoint detection error during the decoding phase of an isolated word recognition task. Since the conventional Viterbi algorithm explores only the search space whose boundaries are fixed to the endpoints of the segmented utterance by the endpoint detector, the recognition performance is highly dependent on the accuracy level of endpoint detection. Inaccurately segmented word boundaries lead directly to recognition error. In order to relax the degradation of recognition accuracy due to endpoint detection error, we describe an unconstrained search of word boundaries and present an algorithm to explore the search space with efficiency. The proposed algorithm was evaluated by performing a variety of simulated endpoint detection error cases on an isolated word recognition task. The proposed algorithm reduced the Word Error Rate (WER) considerably, from 84.4% to 10.6%, while consuming only a little more computation power.

Keywords: *Noise robust recognition, Word boundary detection, Viterbi decoding*

I. Introduction

From a practical point of view, one of the most critical factors that affects recognition performance of an isolated word recognition task is the level of endpoint detection accuracy. This is because the boundaries of the search space to be explored to find the most probable word for the given acoustic observations are fixed to the endpoints of the segmented utterance by so-called endpoint constraint. Therefore, inaccurately endpoint-detected utterances cannot help leading to recognition errors. There have been many approaches to address the endpoint detection problem. These approaches can be classified into two categories: one tries to segment word boundaries as accurately as possible for variations in the surrounding environment by using noise adaptation techniques and/or statistically different information between speech and the non-speech signal [1][2], the other uses acoustic filler models to absorb non-speech signals under a keyword-spotting framework [3][4]. Both methods have in common that prior knowledge of the noise or non-speech signal is required to classify the non-speech signal. In other words, detection accuracy can deteriorate when unexpected noise sources occur. In this paper, we will present a different approach. The goal of the proposed method is to explore the word boundary unconstrained search space to compensate for endpoint detection errors without any prior information about the noise or non-speech signal. The remainder of this paper is organized as follows: In Section II, we describe the basic idea of the word boundary unconstrained search to compensate for endpoint detection error in the decoding phase. In Section III, the conventional Viterbi decoding algorithm is reviewed briefly. In Section IV, we present the modified Viterbi algorithm, which explores the search space with efficiency

Corresponding author: Hoon Chung (hchung@etri.re.kr)
ETRI, 161 Gajeong-Dong Yuseong-Gu, Daejeon 305-350, Korea

in a time-synchronous fashion. In the last Section, we present recognition experiments performed on simulated endpoint detection error conditions.

II. The Word Boundary Unconstrained Search

The basic idea of word boundary unconstrained search to compensate for endpoint detection error is simple. Since it cannot be guaranteed that word boundaries are segmented accurately by the endpoint detector, instead of fixing the boundaries of the search space to the endpoints of a segmented utterance we assume that the correct word boundaries should be within the predefined start and end boundary margins and try to explore the search space iteratively by varying the word boundaries. Fig. 1 illustrates an example of the word boundary unconstrained search process.

In Fig. 1, NS represents the non-speech signal detected as speech due to endpoint detection error. Assuming that the start boundary margin is the first 6 frames and the end boundary margin is the last 8 frames, there is a total number of 48 possible word boundaries; the same number of Viterbi decoding algorithms should be performed on these distinct segments to get the best word, which produces maximum a posterior (MAP) probability normalized with the length of each segment. Even though this exhaustive search process works well, as expected, for adverse endpoint detection conditions, in practice it is hard to use due to the huge computational needs and frame-asynchronous characteristics. So, we present a modified algorithm, which explores the word boundary unconstrained search space very efficiently in a frame-synchronous manner.

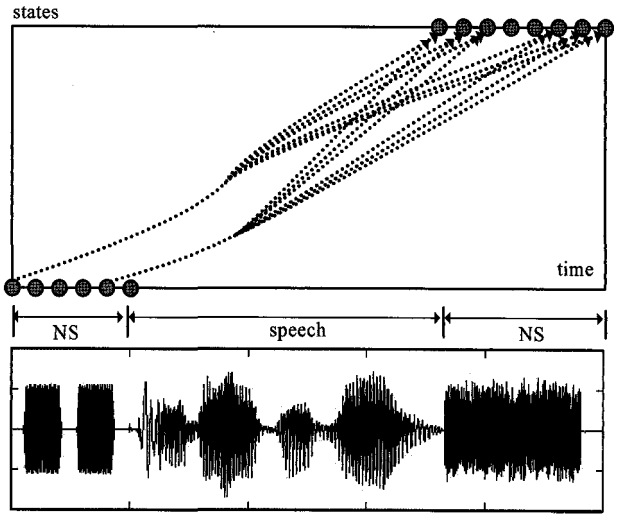

Fig. 1. An example of word boundary unconstrained search for an inaccurately segmented utterance.

III. The Viterbi Algorithm

In this section, we briefly discuss the conventional Viterbi algorithm and how the conventional Viterbi algorithm can be converted to explore the word boundary unconstrained search space. The Viterbi algorithm is a DP algorithm, which finds the optimal state sequence that maximizes a posterior probability for a given Hidden Markov Model (HMM) $\lambda=\{\pi,A,B\}$ and acoustic observations $X=\{x_1,x_2,...,x_T\}$ by defining a variable $\delta^t(i)$ [5].

$$\delta^t(i)=\max_{q_1,q_2,...,q_{t-1}} P(q_1,q_2,...,q_{t-1},q_t=i,x_1,x_2,...,x_{t-1},x_t\mid\lambda)$$
$$=\text{conditional probability that the word }\lambda\text{ produces the acoustic observations }x_1,x_2,...,x_{t-1},x_t \quad (1)$$

The optimal state likelihood is then calculated by the Viterbi algorithm as

1. Initialization:
$$\delta^1(i)=\pi_i\cdot b_i(x_1),\ 1\le i\le N$$
2. Recursion:
$$\delta^t(j)=\max_i\{\delta^{t-1}(i)\cdot a_{ij}\}\cdot b_j(x_t),\ 1\le i,j\le N, 2\le t\le T$$
3. Termination:
$$P^*=\arg\max_i\{\delta^T(i)\} \quad (2)$$

As expressed in (2), the initialization and termination steps limit the boundaries of the search space to the first frame $t=1$ and the last frame $t=T$. By relaxing this endpoint constraint, we can achieve word boundary unconstrained search.

IV. The Modified Viterbi algorithm

As a result of relaxing the endpoint constraint, there

should be the number of hypothetical word boundaries, the product of the start boundary margins and the end boundary margins, and it takes time nearly proportional to the number of possible word boundaries to explore the search space iteratively with the conventional Viterbi algorithm. Hence, in order to reduce the computational loads of this exhaustive search with the conventional Viterbi algorithm, we will describe some modifications to the conventional Viterbi algorithm in this Section.

4.1. Start Point Unconstraint

As depicted in Fig. 2, there is the same number of partial hypotheses arriving to state i at time t as the number of start boundary margins. Each hypothesis can be expressed in terms of the variable introduced in the time conditioned approach [6][7].

$$\delta_\tau^t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P(q_1,q_2,\cdots,q_{t-1},q_t=i,x_\tau,x_{\tau+1},\cdots,x_{t-1},x_t \mid \lambda) \quad (3)$$

where $\delta_\tau^t(i)$ denotes conditional probability that a given HMM λ produces the partial acoustic observation sx_τ^t that starts from time τ and ends at time t. In a maximum approximation point of view, there is only one hypothesis arriving to state$i$ at time t. Hence, (1) can be expressed in terms of $\delta_\tau^t(i)$ if $\delta_\tau^t(i)$ is properly normalized with respect to time.

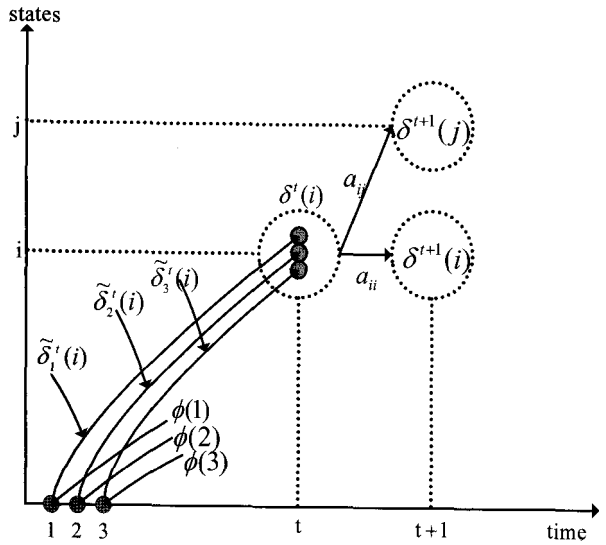

Fig. 2. Partial hypotheses arriving to state i at time t starting from different start points.

$$\delta^t(i) = \max_\tau \{\tilde{\delta}_\tau^t(i)\}, \quad \tilde{\delta}_\tau^t(i) = \phi(\tau)\cdot\delta_\tau^t(i) \quad (4)$$

where $\tilde{\delta}_\tau^t(i)$ is the normalized likelihood as if it started from time $\tau=1$ by normalization weight $\phi(\tau)$.

Since Viterbi decoding is the process to find the optimal state sequence, it is reasonable to make the assumption that a newly starting hypothesis with state i from time τ has made the transition from the state with the maximum likelihood at the previous time $\tau-1$. We can define the normalization weight $\phi(\tau)$ as follows:

$$\phi(\tau) = \begin{cases} 1.0, & \tau = 1 \\ \max\limits_i \left(\delta^{\tau-1}(i)\right), & 1 \le i \le N,\ 2 \le \tau \le D_b \end{cases} \quad (5)$$

where D_b denotes the start boundary margin and $\phi(\tau)$ means the maximum likelihood at time $\tau-1$. Since most speech recognizers use the beam pruning technique to kill unlikely hypotheses compared to the most probable hypothesis, the normalization weight $\phi(\tau)$ can be obtained without more computational load and the normalization can be performed in time-synchronous fashion.

4.2. End Point Unconstraint

The endpoint unconstraint can be achieved by extending the termination region from $t=T$ to endpoint boundary margin $T-D_e \le t \le T$. Similar to the unconstrained start point case, there are many terminating hypotheses representing different lengths of acoustic observations, and we normalize the likelihood score as follows:

$$P^* = \max_i \left\{(\delta^t(i))^{1/t}\right\} \quad (6)$$

Unlike the start point unconstraint case, the normalization in the termination step does not affect time-synchronous processing. So, we can simply normalize the likelihood scores with respect to time. Fig. 3 illustrates the normalization process and weights.

In Fig. 3, newly starting hypothesis from time τ is normalized by $\phi(\tau)$ and terminating hypothesis is normalized with respect to time.

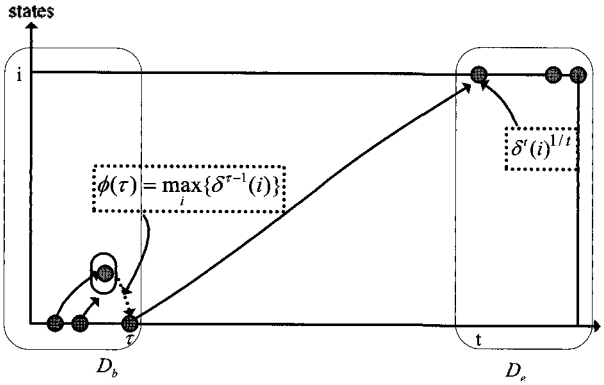

Fig. 3. Normalization weights for the modified Viterbi decoding algorithm.

4.3. Modified Viterbi Algorithm

We have modified the initialization and termination steps of the conventional Viterbi algorithm to accomplish unconstrained word boundary search efficiently. By replacing these two steps with the modified ones and inducing the recursion step as follows, we can obtain the modified Viterbi algorithm.

1.*Initialization* :

$$\tilde{\delta}_\tau^\tau(i) = \pi_i \cdot \phi(\tau) \cdot b_i(x_\tau),\ 1 \le i \le N,\ 1 \le \tau \le D_b$$

2.Recursion :

$$\tilde{\delta}_\tau^t(j) = \max_i \left\{ \tilde{\delta}_\tau^t(i) \cdot a_{ij} \right\} \cdot b_j(x_t),\ 2 \le t \le T,\ 1 \le i, j \le N$$

$$\begin{aligned} \delta^t(j) &= \max_\tau \left\{ \tilde{\delta}_\tau^t(j) \right\} \\ &= \max_\tau \left\{ \max_i \left(\tilde{\delta}_\tau^{t-1}(i) \cdot a_{ij} \right) \cdot b_j(x_t) \right\} \\ &= \max_i \left\{ \max_\tau \left\{ \tilde{\delta}_\tau^{t-1}(i) \right\} \cdot a_{ij} \right\} \cdot b_j(x_t) \\ &= \max_i \left\{ \delta^{t-1}(i) \cdot a_{ij} \right\} \cdot b_j(x_t) \end{aligned}$$

3.Termination : (7)

$$P^* = \max_i \left\{ \delta^t(i)^{1/t} \right\},\ T - D_e \le t \le T$$

where D_b and D_e denote word boundary margins within which we assume that correct word boundaries exist. As can be seen in (7), the conventional Viterbi algorithm can be converted to the proposed Viterbi algorithm with minor modifications in the initialization and termination steps.

V. Experiment And Results

In order to evaluate the performance of the proposed Viterbi decoding algorithm in a variety of endpoint detection error conditions, we defined a function that simulates these cases.

$$\begin{aligned} &epd(t, n_b, d_b, p_b, n_e, d_e, p_e, S) = \\ &n_b(t) \cdot \mathrm{Rect}(t, 1, d_b) + w(t) \cdot \mathrm{Rect}(t, d_b, p_b) + S(t) \cdot \mathrm{Rect}(t, d_b + p_b, T) + \\ &w(t) \cdot \mathrm{Rect}(t, d_b + p_b + T, p_e) + n_e(t) \cdot \mathrm{Rect}(t, d_b + p_b + T + p_e, d_e), \\ &\mathrm{Rect}(t, s, d) = \begin{cases} 1, s \le t \le s + d \\ 0, others \end{cases} \end{aligned} \quad (8)$$

where $n_b(t)$ and $n_e(t)$ denote the non-speech signal, d_b and d_e are the non-speech signal durations, p_b and p_e are pause durations, $w(t)$ is white Gaussian noise with zero mean and unit variance, and $S(t)$ denotes an accurately segmented utterance. Fig. 4 depicts the meaning of the variables of function $epd(t, n_b, d_b, p_b, n_e, d_e, p_e, S)$.

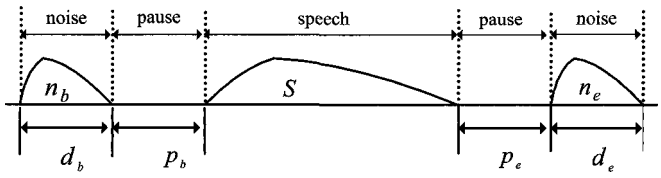

Fig. 4. General example of an inaccurately endpoint-detected utterance.

We prepared ten kinds of non-speech signals to simulate endpoint detection error cases caused by various non-speech signals such as musical noise, barking sounds, the ringing of a telephone, a baby crying, the sound of a door closing, laughing sounds, keyboard noise, wind noise, the sound of an air conditioner, and water sounds. We allowed the non-speech signal to last from 0 ms to 1000 ms and the pause silence from 100 ms to 500 ms. The proposed algorithm was tested on an isolated word recognition task in which the active vocabulary consisted of 1130 phonetically balanced Korean words and correctly segmented 2260 utterances composed of two sets spoken by twenty-one people. We generated 2260 corresponding endpoint detection error utterances using the $epd(t, n_b, d_b, p_b, n_e, d_e, p_e, S)$ function where all variables except speech S are assumed to have equal occurrence distribution. We used the tied-state triphone model of 1860 states in which each state was represented by a Gaussian mixture comprised of 12 Gaussians components. We extracted MFCCs, C0 energy, and their delta, leading to 26 features. The experiment was performed by varying word boundary margins D_b and D_e as below.

$$D_b = T \cdot \text{Word boundary margin ratio}$$
$$D_e = T - (T \cdot \text{Word boundary margin ratio}) \quad (9)$$

where T is the total frame number of a utterance. Fig. 5 shows the Word Error Rate (WER) of the inaccurately endpoint-detected (EPD) utterances as well as the accurately segmented utterances. It can be seen that the modified Viterbi algorithm reduces the WER of the inaccurately endpoint-detected utterances considerably. The WER is reduced from 84.39% to 10.6% when extending the word boundaries margin by 30% for both sides, while the recognition accuracy degrades very little for the accurately segmented utterances. This means that the recognition performance for the accurately segmented utterance is little affected by adopting the proposed Viterbi algorithm.

Fig. 6 shows the additional computation loads required by the proposed Viterbi algorithm. The computation is measured by a real-time factor. It is defined as the division of the total recognition time by the total time of the speech utterances on a 2.7GHz Pentium 4 workstation.

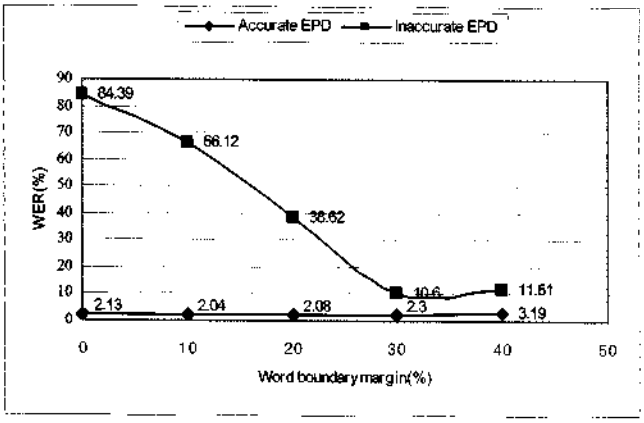

Fig. 5. Recognition performance of the proposed Viterbi algorithm.

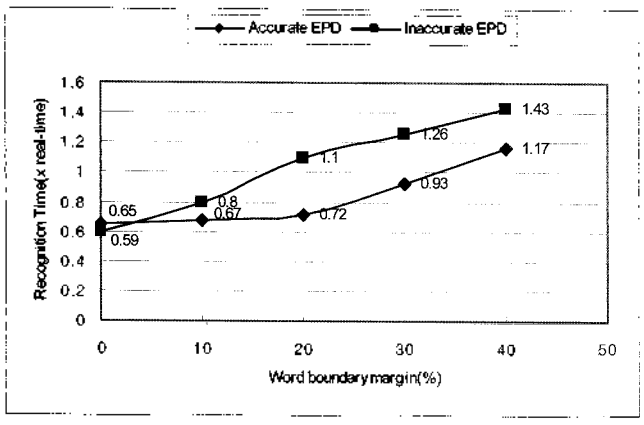

Fig. 6. Additional computation load of the proposed Viterbi algorithm.

As can be seen in Fig. 6, the modified Viterbi algorithm takes computation time nearly proportional to the word boundary margin. This computational load is relatively low in comparison with the exhaustive search using the conventional Viterbi algorithm. If we perform the conventional Viterbi algorithm iteratively for all assumed endpoints of a given word boundary margin, it takes time proportional to the square of the number of word boundary margin frames. In the mean while, it takes more time to explore the word boundary unconstrained search space of an inaccurately endpoint-detected utterance than the search space of an accurately endpoint-detected utterance for the same word boundary margin. This is because non-speech signals make more sub-word models to be survived.

VI. Conclusion

In this paper, we describe word boundary unconstrained search to compensate for endpoint detection error on isolated word recognition and present its efficient implementation. The modified Viterbi algorithm achieved considerable reduction of the WER (from 84.4% to 10.6%) in a variety of simulated endpoint detection error cases while maintaining almost the same level of accuracy on the accurately segmented utterances. The conventional Viterbi algorithm can be easily converted to the proposed algorithm with minor modifications for the initialization and termination steps. In a sense, the conventional Viterbi algorithm can be regarded as a special case of the proposed algorithm where the word boundary margin is fixed to the endpoints of an utterance.

References

1. Chin-Teng Lin, Jiann-Yow Lin and Gin-Der Wu, "A robust word boundary detection algorithm for variable noise-level environment in cars," IEEE Transactions on Intelligent Transportation Systems, 3, 89-101, March 2002
2. S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," IEEE Trans. Speech Audio Processing, 8, 478-482, July 2000.
3. R. El Meliani and D. O'Shaughnessy, "New efficient fillers for unlimited word recognition and keyword spotting," ICSLP, 590-593, Oct. 1996
4. C. Tschope, D. Hentschel, M Wolff, M. Eichner and R. Hoffmann, "Classification of non-speech acoustic signals using structure models," IEEE ICASSP, 653-656, May 2004

5. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, (NJ: Prentice-Hall, 1993), pp. 339-340
6. S. Ortmanns, H. Ney, F. Seide, and I. Lindam, "A comparison of time conditioned and word conditioned search techniques for large vocabulary speech recognition," ICSLP, 2091-2094, Oct. 1996
7. S. Ortmanns and H. Ney, "The time-conditioned approach in dynamic programming search for LVCSR," **IEEE** Trans. Speech Audio Processing, 8, 676-687, Nov. 2000.

[Profile]

• **Hoon Chung**

He received the B.S. and M.S. degrees in Electronics Engineering from Kangwon National University, Korea in 1994 and 1996, respectively. Since 2004, he has been working for Spoken Language Processing Team of ETRI, Daejeon, Korea. His research interests include very large vocabulary speech recognition and fast search algorithm.

• **Ikjoo Chung**

He received the B.S., M.S and Ph. D degrees in Electronics Engineering from Seoul National University, Korea in 1986, 1988 and 1992, respectively. Since 1992, he has been with Kangwon National University and he is now a professor at the Dept. of Electrical and Electronic Engineering, College of Information Technology. His research interests include embedded speech recognition and real-time implementation.