Word Boundary Unconstraint Viterbi Search For Robust Speech Recognition

Hoon Chung, SungJoo Lee and YunKeun Lee

Electronics and Telecommunications Research Institute, Daejeon, Korea
E-mail: {hchung, lee1862, yklee}@etri.re.kr Tel: +82-860-5742

Abstract — *In this paper, we propose a word boundary unconstraint Viterbi algorithm for robust speech recognition against endpoint detection error on isolated word task domain. In general, one-keyword spotting framework is used to achieve robust recognition by absorbing non-speech events with acoustic filler models. One drawback of such an approach is that there is little improvement or it can even hurt performance if unexpected non-speech events occur, whose spectral characteristics are not trained into acoustic filler models. However, it is unrealistic to prepare acoustic filler models considering all kinds of non-speech events especially occurring in mobile environment. Therefore, we propose another approach where non-speech events are absorbed implicitly by relaxing endpoints constraint of Viterbi algorithm. The experimental results show that the algorithm reduces the word error rate from 80.2% to 10.6% for inaccurately endpoint-detected utterances while consuming a little more computation.*

Index Terms — Keyword-spotting, acoustic filler model, Viterbi decoding

I. INTRODUCTION

Even though current state-of-the-art speech recognition is large vocabulary continuous speech recognition (LVCSR) system which can recognize hundreds of thousands of words spoken naturally, isolated word recognition is still used in many commercial areas, especially in mobile environments. For example, name dialing in a cellular phone or command-and-control system in a car. In order for speech recognition to survive in those commercial areas, most of all, it must work robust irrespective of surrounding noise conditions. So, a lot of approaches have been proposed to deal with noise robustness problem. Among them, in this paper, we focus on robust recognition on endpoint detection errors. Endpoint detection is a frontend process to segment a portion which is assumed to be speech from an input signal. In isolated word recognition, the accuracy level of endpoint detection highly affects the total recognition performance. Therefore, various methods have been proposed to achieve robust recognition against endpoint detection error, and these approaches can be broadly classified into one of two categories: one in pre-processing stage and the other in decoding stage. The methods in pre-processing stage try 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]. While, the methods in decoding stage make use of one-keyword spotting framework to absorb non-speech events with acoustic filler models [3][4]. In ideal, acoustic filler models are trained to capture characteristics of every type of non-speech events that can occur in real situation. However, it is hard to prepare filler models that consider all types of non-speech events, and this mismatch between acoustic fillers and actual non-speech events may degrade the recognition performance. Therefore, in this paper, we propose another approach which does not use any of explicit acoustic filler models but performs the same function of one-keyword spotting by introducing a word boundary unconstrained Viterbi algorithm.

The remainder of this paper is organized as follows: In Section II, we give a brief overview of one-keyword spotting. In Section III, we explain the idea of word boundary unconstrained search. In Section IV, we present a word boundary unconstraint Viterbi algorithm. In the last Section, we present recognition experiments performed on simulated endpoint detection error conditions.

II. ONE-KEYWORD SPOTTING

One-keyword spotting is a process to find an optimal word $\hat{w}$ for an input observation X, which satisfies the following condition:

$$\begin{aligned}\hat{W} &= \arg\max_{w} P(X \mid W) \\ &= \arg\max_{\tau,t,w} \tilde{P}(X_{\tau}^{t} \mid W) \\ &= \arg\max_{\tau,t,w} \left\{ P(X_{1}^{\tau-1} \mid NS) \cdot P(X_{\tau}^{t} \mid W) \cdot P(X_{t+1}^{T} \mid NS) \right\}\end{aligned} \quad (1)$$

where $P(X_1^{\tau-1}|NS)$ and $P(X_{t+1}^{T}|NS)$ denote the conditional probabilities that the acoustic filler model NS produces partial observations for $X_1^{\tau-1}$ and X_{t+1}^{T} individually and $P(X_{\tau}^{t}|W)$ denotes the conditional probability that a word model produces the observations starting from time τ and ending at time t. In general, one-keyword spotting is configured as depicted in Fig. 1

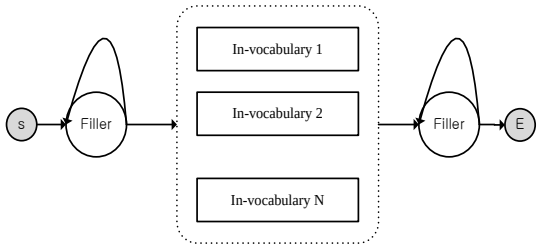

Fig. 1: A typical finite state network for one-keyword spotting.

where the filler models are configured to self loop for absorbing arbitrary sequences of non-keyword segments including non-speech signal. In HMM-based approach, each sub-word model is trained with a HMM $\lambda = \{\pi, A, B\}$, and keyword models are constructed by concatenating sub-word HMMs according to their pronunciations. In decoding stage, Viterbi algorithm is performed to find the optimal state sequence that maximizes a posterior probability for a given HMM and an input observation X by defining a variable $\delta^t(i)$ [5].

$$\delta^t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P(q_1, q_2, \ldots, q_{t-1}, q_t = i, x_1, x_2, \ldots, x_{t-1}, x_t \mid \lambda)$$
$$= \text{conditional probability that the word } \lambda \text{ produces} \quad (2)$$
$$\text{the acoustic observations } x_1^t$$

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

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

In one-keyword spotting, the conditional probabilities which observe the non-speech segments are defined in terms of $P(X_1^{\tau-1}|NS)$ and $P(X_{t+1}^T|NS)$. In this paper, we describe a method to estimate these probabilities implicitly using the word boundary unconstrained search strategy.

III. WORD BOUNDARY UNCONSTRAINT SEARCH

The fundamental idea of the word boundary unconstraint search is as follows: we make assumption that correct word boundaries exist with a predefined margin for the automatically marked boundaries by endpoint detection, and then explore the search space exhaustively by varying the word boundaries. Fig. 2 illustrates an example of the word boundary unconstrained search.

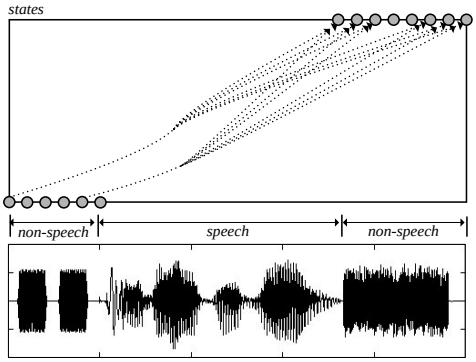

Fig. 2: An example of word boundary unconstrained search for a segmented utterance containing non-speech events.

Assuming that the start boundary margin is 6 frames and the end boundary margin is 8 frames, there are 48 pairs of word boundaries and then we have to run 48 times of Viterbi decoding to find an optimal state sequence. Even though this exhaustive search process works well for adverse endpoint detection conditions, in real applications, it is unpractical due to the huge computational needs and frame-asynchronous characteristics. So, we present a modified algorithm, which explores the word boundary unconstrained search space efficiently in frame-synchronous manner.

A. Start Point Unconstraint

As depicted in Fig. 3, there is the same number of partial hypothesis which arrives to state i at time t started from time τ as the number of start boundary margin. Each hypothesis can be expressed in terms of the variable introduced in time-conditioned approach [6][7].

$$\delta_\tau^t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P(q_1, q_2, \ldots, q_{t-1}, q_t = i, x_\tau, x_{\tau+1}, \ldots, x_{t-1}, x_t \mid \lambda) \quad (4)$$

where $\delta_\tau^t(i)$ denotes conditional probability that a given HMM λ produces the partial acoustic observations X_τ^t which 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, (2) can be expressed in terms of $\delta_\tau^t(i)$ if $\delta_\tau^t(i)$ is properly normalized with respect to time.

$$\delta^t(i) = \max_\tau \{\tilde{\delta}_\tau^t(i)\}, \quad \tilde{\delta}_\tau^t(i) = \phi_1^\tau \cdot \delta_\tau^t(i) \quad (5)$$

where $\tilde{\delta}_\tau^t(i)$ is normalized likelihood as if it started from time $\tau = 1$ by normalization weight ϕ_1^τ.

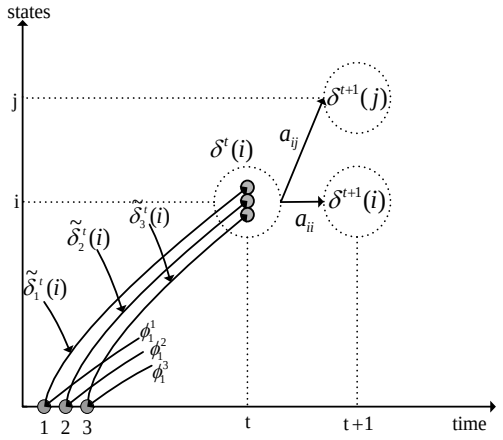

Fig. 3: Partial hypotheses arriving to state i at time t starting from different start points

Since the Viterbi decoding is the process to find the optimal state sequence, it is reasonable to make assumption that newly starting hypothesis with state i from time τ has been made transition from the state with the maximum likelihood at the previous time $\tau - 1$ and we can define the normalization weight ϕ_1^τ as follows.

$$\phi_1^\tau \approx \begin{cases} 1.0, & \tau = 1 \\ \max_i \left(\delta^{\tau-1}(i) \right), & 1 \le i \le N,\ 2 \le \tau \le D_b \end{cases} \quad (6)$$

where D_b denotes the start boundary margin and ϕ_1^τ means the maximum likelihood at time $\tau - 1$. Since most speech recognizer uses beam pruning technique to kill unlikely hypotheses compared to the most probable hypothesis, the normalization weight ϕ_1^τ can be obtained without more computational load and the normalization can be performed in time-synchronous fashion.

B. End Point Unconstraint

As can be seen in Fig. 4, there are also many terminating hypotheses representing different length of acoustic observations as a result of endpoint unconstraint.

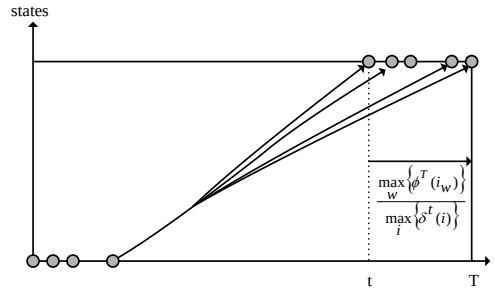

Fig. 4: Terminal hypotheses arriving to state i at time t.

So, we have to normalize the likelihood score as if it terminated at the final time T with normalization weight ϕ_t^T as follows:

$$P^* = \delta^t(i) \cdot \phi_t^T, \quad \phi_t^T = \frac{\max_w \left\{\phi^T(i_w)\right\}}{\max_i \left\{\delta^t(i)\right\}} \tag{7}$$

where i_w denotes the terminal state of the word W, this normalization concept is identical to that of start point unconstraint case.

IV. WORD BOUNDARY UNCONSTRAINT VITERBI

We have modified the initialization and termination steps of the Viterbi algorithm to accomplish unconstrained word boundary search. 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 :

$$\begin{aligned}
\tilde{\delta}_\tau^t(j) &= \max_i \left(\tilde{\delta}_\tau^t(i) \cdot a_{ij}\right) \cdot b_j(x_t), \ 2 \le t \le T, \ 1 \le i, j \le N \\
\delta^t(j) &= \max_\tau \left\{\tilde{\delta}_\tau^t(j)\right\} \\
&= \max_\tau \left\{\max_i \left(\tilde{\delta}_\tau^{t-1}(i) \cdot a_{ij}\right) \cdot b_j(x_t)\right\} \\
&= \max_i \left\{\max_\tau \left\{\tilde{\delta}_\tau^{t-1}(i)\right\} \cdot a_{ij}\right\} \cdot b_j(x_t) \\
&= \max_i \left\{\delta^{t-1}(i) \cdot a_{ij}\right\} \cdot b_j(x_t)
\end{aligned} \tag{8}$$

3.Termination :

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

where D_b and D_e denote word boundary margins within which we assume that correct word boundaries exist. As can be seen in (8), the conventional Viterbi algorithm can be converted to the proposed Viterbi algorithm with minor modifications in the initialization and termination steps.

V. EXPERIMENTS AND RESULTS

A. Korean PBW Test Domain

The proposed Viterbi algorithm is tested on a phonetically balanced word (PBW) domain, which is an isolated word recognition domain composed of a training corpus with 90400 utterances (400 males, 400 females), a vocabulary of 1130 words and 2260 test utterances. Speech signal is sampled at 16kHz, and the frame length is 20ms with 10ms shift. Each speech frame is parameterized as a 26-dimensional feature vector containing 12 MFCCs, C0 energy and their delta feature. We trained tied-state triphone models of 1860 states in which each state was represented by a Gaussian mixture model (GMM) comprised of 12 Gaussian components.

B. Test Set Preparation

For evaluation, we prepare two sets of test utterances from the 2260 utterances. One is a clean test set for simulating correctly endpoint detected cases, and the other is a corrupted test set for simulating incorrectly endpoint detected cases. The clean test set is prepared by segmenting speech portions from the 2260 utterances manually but the corrupted test set is prepared by concatenating noise and speech artificially.

To generate corrupted test set considering various noise conditions, we prepared 10 kinds of noises such as, a horn, general babble, car, motor cycle, barking, a baby crying, a door closing, an electric fan, phone ring, and water. We call the first five noises as "outdoor" class noise and rest of them as "indoor" class noise, and two acoustic filler models are trained with noises in that class. Corrupted set is prepared with $epd(t, n_b, d_b, p_b, n_e, d_e, p_b, s)$ function defined as follows:

$$\begin{aligned}
&epd(t, n_b, d_b, p_b, n_e, d_e, p_b, s) = \\
&n_b(t) \cdot RECT(t, d_b) + w(t) \cdot RECT(t, d_b, p_b) + \\
&s(t) \cdot RECT(t, d_b + p_b, T) + \\
&w(t) \cdot RECT(t, d_b + p_b + T, p_e) + \\
&n_e(t) \cdot RECT(t, d_b + p_b + T + p_e, d_e) \\
&where, RECT(t, s, d) = \begin{cases} 1, & s \le t \le s + d \\ 0, & others \end{cases}
\end{aligned} \tag{9}$$

where $n_b(t)$ and $n_e(t)$ represent one of 10 noise signals, d_b and d_e are durations of individual noise signals, 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 correctly segmented utterance. We allowed noise signals to last from 0ms to 2000 ms and the pause silence from 0ms to 1000ms with uniform probability. Fig. 5 depicts the meaning of $epd(t, n_b, d_b, p_b, n_e, d_e, p_b, s)$ function.

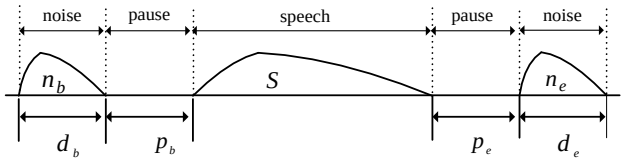

Fig. 5: An example of an inaccurately endpoint-detected utterance.

The experiment was performed by varying word boundary margins D_b and D_e as below.

$$\begin{aligned}
D_b &= T \cdot boundary\ margin\ ratio \\
D_e &= T \cdot (1 - boundary\ margin\ ratio)
\end{aligned} \tag{10}$$

where T is the total number of frames.

C. Baseline Performance

We took experiments to see the baseline performances on various conditions: performance of isolated word recognition system for correctly and incorrectly endpoint detected utterances, performance degradation of one-keyword spotting for unexpected noises cases [3],

and performance of one-keyword spotting for expected noises cases. Table 1 shows word error rates (WERs) for these cases.

TABLE I
WORD ERROR RATES FOR BASELINE CONDITIONS

Baseline system	Test set	WER (%)
ISO	Clean	2.13
ISO	Corrupted	84.4
KWS with "outdoor filler"	Corrupted	52.4
KWS with "indoor filler"	Corrupted	48.8
KWS with "out & in fillers"	Corrupted	7.2

ISO: Isolated recognition, KWS : keyword spotting

In Table 1, the 1st and 2nd rows show that inaccurate endpoint detection seriously degrades recognition performance, the 3rd and 4th rows show that unexpected noise events also degrades recognition performance and the 5th row shows that it improves performance to use matched filler models.

Fig. 6 shows the results of the word boundary unconstrained search for the corrupted test set as well as the clean test set. It reduces the WER of the corrupted set considerably but it does not hurt accuracy of the clean set. The WER is reduced from 80.2% to 10.6% when extending the word boundary margins by 30% for both sides. It is comparable to that of the one-keyword system which uses matched filler models.

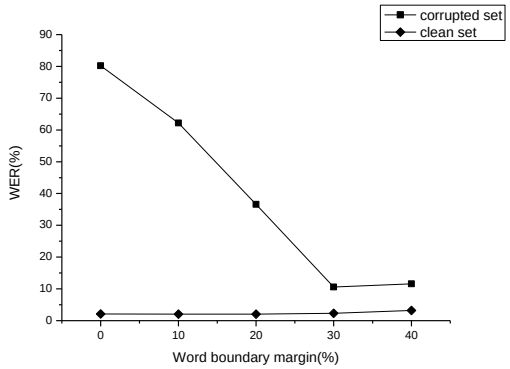

Fig. 6: Recognition performance of the proposed Viterbi algorithm

Fig. 7 shows the computational complexity of the proposed algorithm in real-time factor, which is defined as the division of the total recognition time by the total time of the speech utterances on an embedded device with a fixed-point processor running at 400MHz.

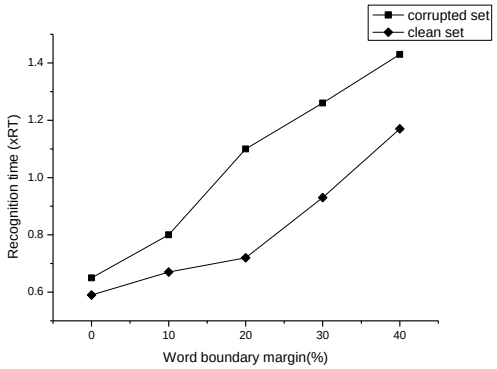

Fig. 7: Computational complexity of the proposed Viterbi algorithm.

The proposed 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 will take time proportional to the square of the number of word boundary margin frames.

VI. CONCLUSION

In this paper, we describe word boundary unconstrained search as an alternative approach for one-keyword spotting to compensate for endpoint detection error on isolated word recognition. The modified Viterbi algorithm achieved a considerable reduction of the WER (from 80.2% to 10.6%) in a variety of simulated endpoint detection error cases without any explicit acoustic filler models. This result is comparable to that of the one-keyword spotting system with the explicit filler models corresponding to non-speech signals. We introduced two weights to normalize the different hypotheses probabilities representing different lengths of speech segments. From a keyword spotting point of view, they find the conditional probabilities producing the non-speech signals and thus they can be regarded as implicit filler model probabilities. Most of all, the proposed algorithm does not require any prior knowledge of non-speech signals, while the conventional keyword-spotting recognizers give poor recognition performance without all kinds of non-speech signals enough to model the acoustic filler accurately. In addition, the proposed algorithm can be implemented easily with minor modifications for the initialization and termination steps of the conventional Viterbi algorithm.

REFERENCES

[1] Chin-Teng Lin; Jiann-Yow Lin; Gin-Der Wu "A robust word boundary detection algorithm for variable noise-level environment in cars**"** Intelligent Transportation Systems, IEEE Transactions on , Volume: 3 , Issue: 1 , March 2002 pp.89.101

[2] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," IEEE Trans. Speech Audio Processing, vol. 8, pp. 478.482, July 2000.

[3] El Meliani, R.; O'Shaughnessy, *D,* **"**New efficient fillers for unlimited word recognition and keyword spotting" Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on , Volume: 2 , pp590.593 Oct. 1996

[4] Tschope. C, Hentschel. D, Wolff. M, Eichner, M, Hoffmann, *R."*Classification of non-speech acoustic signals using structure models" Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on , Volume: 5, pp 653.6 May 2004

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

[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," in Proc. IEEE Int. Conf. Spoken Language Processing, Philadelphia, PA, Oct. 1996, pp. 2091.2094.

[7] S. Ortmanns and H. Ney, "The time-conditioned approach in dynamic programming search for LVCSR," IEEE Trans. Speech Audio Processing, vol. 8, pp. 676.687, Nov. 2000.