

# Weighted Finite State Transducer–Based Endpoint Detection Using Probabilistic Decision Logic

Hoon Chung, Sung Joo Lee, and Yun Keun Lee

**In this paper, we propose the use of data-driven probabilistic utterance-level decision logic to improve Weighted Finite State Transducer (WFST)-based endpoint detection. In general, endpoint detection is dealt with using two cascaded decision processes. The first process is frame-level speech/non-speech classification based on statistical hypothesis testing, and the second process is a heuristic-knowledge-based utterance-level speech boundary decision. To handle these two processes within a unified framework, we propose a WFST-based approach. However, a WFST-based approach has the same limitations as conventional approaches in that the utterance-level decision is based on heuristic knowledge and the decision parameters are tuned sequentially. Therefore, to obtain decision knowledge from a speech corpus and optimize the parameters at the same time, we propose the use of data-driven probabilistic utterance-level decision logic. The proposed method reduces the average detection failure rate by about 14% for various noisy-speech corpora collected for an endpoint detection evaluation.**

**Keywords:** Endpoint detection, speech recognition, Weighted Finite State Transducer.

## I. Introduction

The endpoint detection problem is conventionally dealt with using two cascaded decision processes. The first process is frame-level speech/non-speech classification based on statistical hypothesis testing, and the second process involves utterance-level speech boundary decision based on heuristic knowledge. The overall performance of endpoint detection is determined through these two processes. However, most research activities have focused on improving the frame-level decision performance by developing robust features [1]–[2], feature combinations [3], and modeling approaches [4]–[9], while little attention has been paid to utterance-level decision, integrating both decision processes, or improving both processes at the same time. This is because the statistical approach provides a way to optimize frame-level decision parameters systematically, whereas a heuristic-knowledge-based approach makes it difficult to define and optimize utterance-level decision parameters. In addition, it is also hard to integrate the two different decision processes.

To solve the integration issue, we proposed Weighted Finite State Transducer (WFST)-based endpoint detection in our previous work [10]. In the proposed WFST-based endpoint detection, both the frame-level decision result and utterance-level heuristic knowledge are represented in WFSTs, and the detection state is determined by composition and best-path search operations. The WFST-based approach provides a straightforward way to integrate the two decision processes. However, the proposed WFST-based approach has the same limitations as in conventional approaches in that the utterance-level decision is based on heuristic knowledge and the decision parameters are tuned sequentially.

To solve these problems, we propose the use of data-driven

---

Manuscript received Jan. 28, 2014; revised June 3, 2014; accepted June 24, 2014.

This work was supported by the Industrial Strategic technology development program, 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Korea).

Hoon Chung (corresponding author, hchung@etri.re.kr), Sung Joo Lee (lee1862@etri.re.kr), and Yun Keun Lee (yklee@etri.re.kr) are with the SW-Content Research Laboratory, ETRI, Daejeon, Rep. of Korea.

probabilistic utterance-level decision logic by modeling the quantized speech/non-speech likelihood ratio trajectory from a segmented speech corpus. The remainder of this paper is organized as follows. In Section II, we briefly review conventional endpoint detection and WFST-based endpoint detection. In Section III, we describe how to train probabilistic utterance-level decision logic from a speech corpus in detail. Finally, in Section IV, experimental results of the endpoint detection test corpus are provided.

## II. Background

In this section, we briefly review both conventional endpoint detection and the proposed WFST-based approach.

### 1. Conventional Endpoint Detection

One of the most widely used frame-level decision methods in recent years is likelihood ratio testing (LRT)-based speech/non-speech classification [5].

$$l(x_t) = p(x_t|H_1) / p(x_t | H_0), \quad (1)$$

$$o_t = H_n, \quad (2)$$

where

$$n = u[l(x_t) - \eta] = \begin{cases} 1 & \text{if } l(x_t) \geq \eta, \\ 0 & \text{if } l(x_t) < \eta, \end{cases} \quad (3)$$

where  $x_t$  is a feature vector at frame  $t$ ,  $H_1$  and  $H_0$  are speech and non-speech hypotheses, respectively,  $u(\cdot)$  is a unit step function,  $\eta$  is a decision threshold, and  $o_t$  is a frame-level decision result at frame  $t$ .

For a frame-level decision sequence,  $O^T = o_1, o_2, \dots, o_T$ , an utterance-level decision makes certain whether the decision sequence satisfies the condition to be classified as a speech segment. This condition is usually based on heuristic knowledge, and Fig. 1 shows widely used decision logic [2].

Here, ‘‘Gap’’ is an integer indicating the required number of frames from a detected endpoint to the actual end of the speech. In practice, such an utterance-level decision is implemented on a finite state machine (FSM) as in the following:

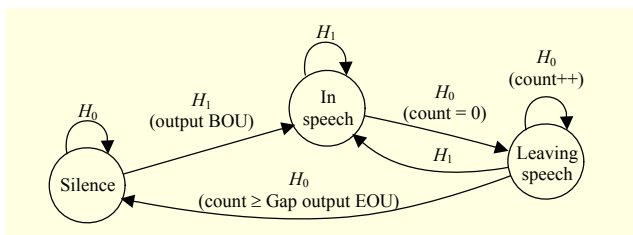


Fig. 1. Utterance-level FSM example.

$$A = (\Sigma, Q, I, \rho, F), \quad (4)$$

where  $\Sigma$  is an input alphabet,  $Q$  is a set of states,  $I$  is an initial state,  $\rho$  is a state transition function, and  $F$  is a set of final states.

### 2. WFST-Based Endpoint Detection

A WFST is an FSM with state transitions labelled with input and output symbols, where each transition has an associated weighting; defined as follows:

$$T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho), \quad (5)$$

where  $\Sigma$  is a finite input alphabet,  $\Delta$  is a finite output alphabet,  $Q$  is a set of states,  $I$  is a set of initial states,  $F$  is a set of final states,  $E$  is a finite set of transitions,  $\lambda$  is the initial weight function, and  $\rho$  is the final weight function [11]–[12].

WFST-based endpoint detection was proposed to deal with two cascaded decision processes under a unified framework, where (1), (2), and (4) are represented in WFSTs, respectively, and the detection state is determined through composition and best-path operations as follows:

$$P = \text{bestpath}(F \circ U), \quad (6)$$

where  $F$  is a frame-level WFST representing equations (1) and (2), and where  $U$  is an utterance-level WFST representing equation (4) with two additional output symbols to mark the begin-of-utterance (BOU) and end-of-utterance (EOU). The endpoint is detected if the output symbol of the last transition of the best path  $P$ ,  $o(e_i)$ , satisfies the following condition:

$$o(e_i) = EOU. \quad (7)$$

As formulated in (6), the performance of the WFST-based endpoint detection is determined by the two WFSTs. Therefore, assuming that the frame-level decision model is  $\theta_F$  and the utterance-level decision model is  $\theta_U$ , the WFST-based endpoint detection model  $\theta$  is defined as follows:

$$\theta = \left[ \underbrace{x_t, p(x_t|H_1), p(x_t|H_0), u(\cdot), \eta}_{\theta_F}, \underbrace{\Sigma, \Delta, Q, I, F, E, \lambda, \rho}_{\theta_U} \right]. \quad (8)$$

As previously stated, there are many research activities on  $x_t$ ,  $p(x_t|H_1)$ , and  $p(x_t|H_0)$ . In this paper, we focus on the defining and optimizing of the parameters  $u(\cdot)$ ,  $\eta$ , and  $\theta_U$ .

## III. Probabilistic Utterance-Level Decision Logic

The fundamental idea of this work is that we handle a frame-level decision as multi-level quantization and an utterance-level decision as a symbol matching process to detect a pre-defined symbol sequence. Figure 2 illustrates an endpoint detection

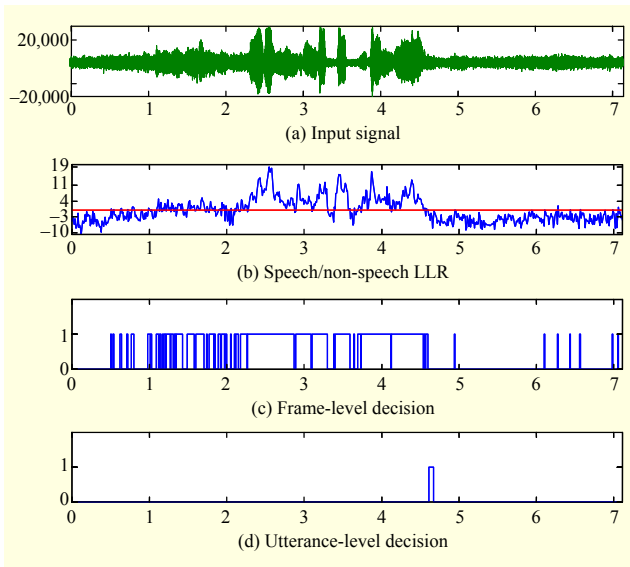


Fig. 2. Endpoint detection process: (a) an input signal, (b) speech/non-speech LLR, (c) frame-level decision results, and (d) utterance-level decision result.

process where (a) is an input signal collected at a bus stop and (b) is a speech/non-speech log-likelihood ratio (LLR) for (a). In a conventional approach, (c) is a frame-level speech/non-speech binary decision sequence and (d) is an utterance-level decision result, where an output of “1” means that an input frame-level decision sequence traverses to an EDU state in an FSM, such as depicted in Fig. 1. However, in the proposed approach, the two decision processes are treated differently. First, the frame-level decision sequence shown in Fig. 2(c) is regarded as an output of 1-bit quantization for Fig. 2(b). Second, the utterance-level decision is regarded as a process to detect a discrete symbol sequence if the sequence is a predefined sequence.

### 1. Speech/Non-speech Likelihood Ratio Quantization

In most cases, endpoint detection is considered from the aspect of voice activity detection (VAD). However, VAD focuses on making an accurate frame-level speech/non-speech decision, whereas endpoint detection has to focus on improving the accuracy of an utterance-level speech segment decision. An accurate VAD is necessary to improve the endpoint detection performance, but it is well known that VAD errors are inevitable and that they degrade the utterance-level decision accuracy. In a statistical model-based VAD or frame-level decision, the frame-level decision errors are affected by the parameters  $\theta_F$  in (8). Among them, in this paper, we focus on the binary decision related parameters,  $u(\cdot)$  and  $\eta$ . From a functional point of view, a binary decision is like a 1-bit quantization since it outputs a zero or one for an input

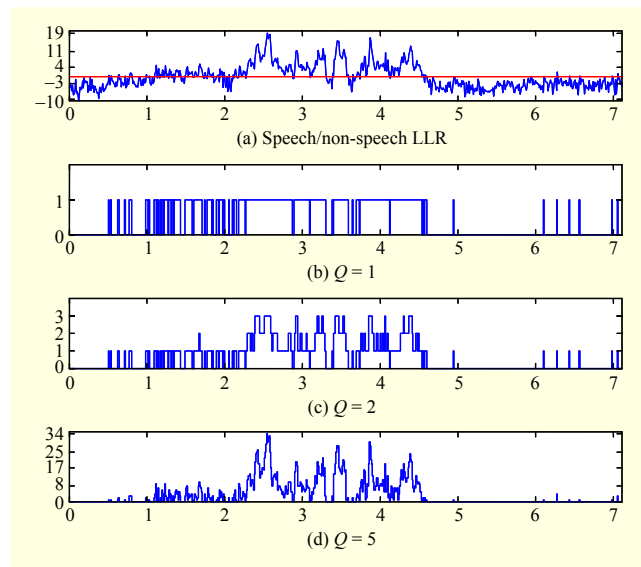


Fig. 3. Speech/non-speech LLR quantization examples: (a) speech/non-speech LLR, (b) 1-bit quantized LLR, (c) 2-bit quantized LLR, and (d) 5-bit quantized LLR.

speech/non-speech likelihood ratio. This means that the propagated frame-level decision errors can be controlled by replacing the binary decision with a general  $Q$ -bit quantization scheme as follows:

$$o_t = H_n, \text{ where } n = \begin{cases} \frac{l(x_t) - \eta}{\omega} + 1 & \text{if } l(x_t) \geq \eta, \\ 0 & \text{if } l(x_t) < \eta, \end{cases} \quad (9)$$

and

$$Q = \log_2 \left( \frac{\max(L)}{\omega} \right), \quad (10)$$

where  $\omega$  is the step size,  $\max(L)$  is the maximum value for all  $l(x_t) - \eta$ , and  $\eta$  is the same frame-level decision threshold in (3). The step size  $\omega$  controls the quantization errors. The smaller the step size used, the smaller the quantization errors that occur. Equation (9) assigns  $H_0$  to a frame whose LLR is lower than  $\eta$ ; otherwise,  $H_n$ , where  $n > 0$  according to  $\omega$ , is assigned. Figure 3 shows an example of a quantized speech/non-speech likelihood ratio: (a) is the same speech/non-speech LLR as in Fig. 2(b), (b) is a 1-bit quantized LLR sequence, (c) is a 2-bit quantized LLR sequence, and (d) is a 5-bit quantized LLR sequence. In this section, we compare two items to a conventional approach. First, Figs. 2(c) and 3(b) show the same results, which indicate that a frame-level speech/non-speech binary decision can be implemented using 1-bit quantization. Second, as shown in Figs. 3(b) through 3(d), the more quantization bits that are used, the less quantization errors that occur, which means that propagated frame-level decision errors in an utterance-level decision can be controlled by

varying the quantization size.

## 2. $N$ -gram-based Utterance-Level Decision Knowledge

In conventional endpoint detection, utterance-level decision logic is usually implemented using heuristic knowledge. We have to question whether such a heuristic approach is effective in defining boundary detection knowledge. In addition, in the proposed frame-level  $Q$ -bit quantization, it is more difficult to define heuristic knowledge since speech/non-speech LLRs are represented as a sequence of  $H_n$  instead of two states,  $H_0$  and  $H_1$ . Therefore, to solve such problems, we propose the use of a training scheme in making utterance-level decision logic. To train utterance-level decision logic from a speech corpus, we assume that an utterance-level decision is simply a symbol matcher to detect a predefined symbol sequence for a transmitted symbol sequence through frame-level  $Q$ -bit quantization. For example, assuming that a bit sequence in Fig. 3(b), 3(c), or 3(d) is transmitted to an utterance-level decision, the utterance-level decision has to make sure that the bit sequence is valid. In this work, we implement this idea using the WFST framework. As formulated in (6), for an input symbol sequence  $O^T$  to be detected through WFST-based endpoint detection, an input projection of an utterance-level WFST has to at least include  $O^T$  for a successful composition. In other words, an utterance WFST used to detect a discrete symbol sequence  $O^T$  is one whose input sequence is the same as  $O^T$ . Therefore, assuming that  $U_i$  is a WFST for the  $i$ th quantized speech/non-speech likelihood ratio sequence, an utterance-level WFST can be obtained by minimizing the WFST that combines all  $U_i$  as follows:

$$U = U_B \otimes \left\{ \min \bigoplus_{i=1}^N U_i \right\} \otimes U_E, \quad (11)$$

where  $U_B$  and  $U_E$  are WFSTs used to mark a BOU and EOU, which are composed of a single transition whose input symbol is  $\langle \text{eps} \rangle$  and output symbol is BOU or EOU. For example, there are three quantized speech/non-speech likelihood ratio sequences  $O_1$ ,  $O_2$ , and  $O_3$  as follows:

$$\begin{aligned} O_1 &= \{H_2, H_2, H_1, H_3, H_3\}, \\ O_2 &= \{H_1, H_1, H_2, H_2\}, \\ O_3 &= \{H_1, H_1, H_3\}. \end{aligned} \quad (12)$$

An utterance-level WFST obtained by (11) is depicted in Fig. 4.

In training an utterance-level WFST, we have to consider practical issues such as different trajectory lengths and unseen trajectories. To cope with these problems, we divide each quantized trajectory as a sequence of quantized subtrajectories having the same length  $N$  and represent an utterance-level WFST  $U$  as a closure of quantized subtrajectories  $U = U_S^*$ , where the quantized subtrajectory WFST  $U_S$  is

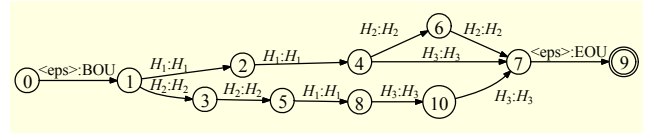


Fig. 4. Utterance-level WFST example.

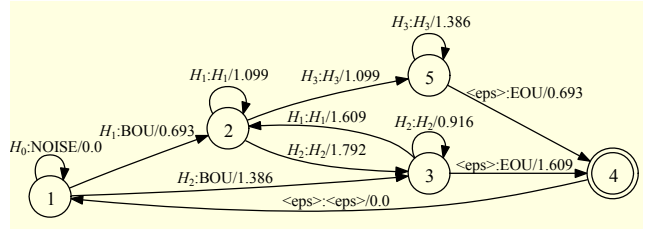


Fig. 5. Example of 2-gram-based utterance-level WFST.

Table 1. Comparison between conventional and proposed approaches.

Decision level	Conventional approach	Proposed approach
Frame	Binary	$Q$ -bit quantization
Utterance	Heuristic FSM	Data-driven $N$ -gram

estimated with  $N$ -gram  $P(o_1, o_2, \dots, o_t)$  as follows:

$$P(o_1, o_2, \dots, o_t) \approx \prod_{i=1}^m P(o_i | o_{i-(N-1)}, \dots, o_{i-1}), \quad (13)$$

where  $o_t$  is the quantized speech/non-speech likelihood ratio at time  $t$ . Figure 5 shows a 2-gram-based utterance-level WFST corresponding to Fig. 4.

In summary, Table 1 lists the differences in the proposed approach as compared to a conventional method. The proposed approach uses  $Q$ -bit quantization at the frame level to control decision errors and  $N$ -gram-based decision logic at the utterance level to represent the trajectory structure in a speech corpus.

The WFST-based endpoint detection in (6) can then be generalized by decision threshold  $\eta$ , step size  $\omega$ , and sub-trajectory length  $N$  as follows:

$$P(\omega, \eta, N) = \text{bestpath}(F(\omega, \eta) \circ U(N)). \quad (14)$$

## IV. Experimental Results

### 1. Endpoint Detection Test Corpus

The proposed approach is evaluated on an in-house endpoint-detection corpus for a Korean voice search. The corpus is composed of 14,000 utterances in total — about 23.5 hours' worth of samples collected from various real-noise

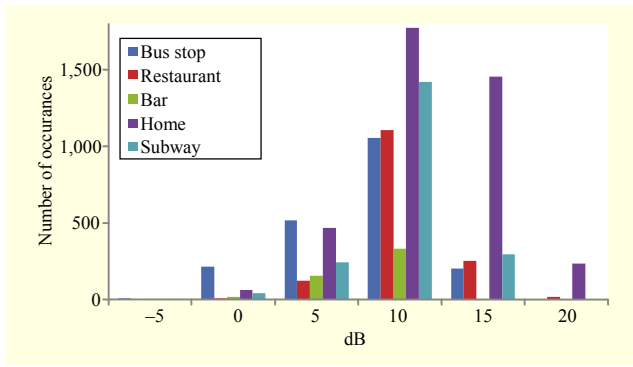


Fig. 6. SNR histogram of each noise condition.

scenarios, such as bus stops, restaurants, bars, homes, and subways, and covering various signal-to-noise ratios (SNRs). Figure 6 displays the SNR histogram of each noise condition.

In this experiment, we used 12,600 utterances to train the speech and non-speech Gaussian mixture models (GMMs) to measure the frame-by-frame speech/non-speech likelihood ratio and utterance-level decision logic; and we used 1,400 utterances for testing. The corpus signal is sampled at 16 kHz, and we used 39-dimensional Mel frequency cepstrum coefficients (MFCCs) composed of 12-dimensional static MFCCs, cepstral energy, and their delta and acceleration as features for speech/non-speech discrimination. In this experiment, we extract feature vectors at every 10 ms for a 20 ms analysis window and use GMMs with 32 components. The performance is measured based on the detection failure rate (DFR), which is defined as follows:

$$DFR = \frac{\#of\ failed\ uttrs}{\#of\ total\ uttrs} \times 100, \quad (15)$$

where the failure count increases if both the beginning and ending utterance points are not within 0.5 s. In actuality, DFR measures any false detection or false rejection errors.

## 2. Baseline Endpoint Detection Performance

In a baseline system evaluation, we use similar utterance-level decision logic as depicted in Fig. 1, and the performance is tuned experimentally by optimizing the frame-level decision threshold, minimum speech frame count, and hang-over frame count. Figure 7 shows the experimental results of the baseline system for a different minimum speech frame count,  $T_m$ , hang-over frame count,  $T_h$ , and different frame-level decision thresholds.

For a baseline system evaluation, we obtained a minimum DFR of 22.07% by setting the frame-level decision threshold to six, the minimum speech frame count to ten frames, and the

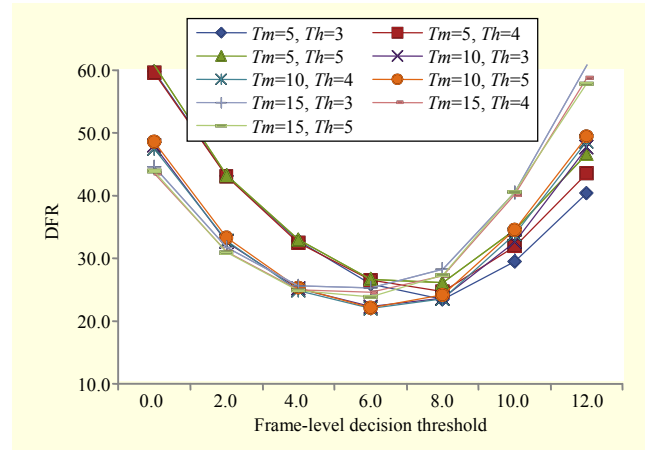


Fig. 7. DFR of baseline system.

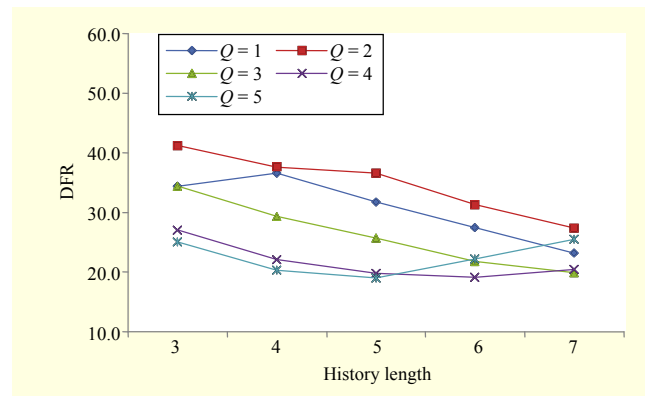


Fig. 8. DFR of proposed approach.

hang-over frame count to four frames.

## 3. Proposed Endpoint Detection Performance

The performance of the proposed approach is measured for quantization bit size  $Q$  and subtrajectory history length  $N$ . Figure 8 shows the experimental results.

As can be seen in Fig. 8, DFR tends to decrease in proportion to both the bit size  $Q$  and the history length  $N$ . This is because  $Q$  is related to a short-term decision error and  $N$  is

Table 2. DFR for noise conditions.

Noise condition	# of Test uttrs	Baseline	Proposed
Bus stop	191	20.42	14.14
Restaurant	173	16.18	9.82
Bar	44	40.91	29.54
Home	386	46.89	37.82
Subway	211	23.22	15.54

Table 3. DFR for SNRs.

SNR	# of Test uttrs	Baseline	Proposed
0	45	42.86	38.10
5	142	35.32	37.00
10	268	31.34	29.48
15	550	16.05	11.19

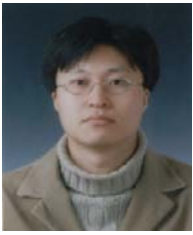
related to a long-term variation. For the proposed approach, we achieve a minimum DFR of 19.0% at  $Q = 5$  and  $N = 5$ . Tables 2 and 3 show the best overall DFRs of the baseline approach and the proposed one for noise conditions and SNRs. The proposed approach outperforms the baseline approach under most noisy conditions.

## V. Conclusion

To improve the overall performance of WFST-based endpoint detection, we proposed the use of probabilistic utterance-level decision logic derived from quantized speech/non-speech likelihood ratio trajectories. In the proposed approach, a frame-level speech/non-speech binary decision is regarded as a quantized likelihood ratio since the final decision is made at the utterance-level. Therefore, to reduce quantization errors, a frame-level binary decision is generalized using  $Q$ -bit quantization, and a heuristic-knowledge-based utterance-level WFST is replaced with an  $N$ -gram-based quantized subtrajectory model to represent quantized discrete symbol sequences. Under the proposed approach, decision-related parameters are optimized from a speech segmented corpus to maximize the overall performance. The experimental results show that the proposed method reduces the failure rate by about 14%. For our future work, we plan to use a discriminative training scheme in designing the frame-level quantization and training the utterance-level decision so as to reduce error rate.

## References

- [1] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-Term Spectro-Temporal and Static Harmonic Features for Voice Activity Detection," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, Oct. 2010, pp. 834–844.
- [2] S.J. Lee et al., "Intra- and Inter-frame Features for Automatic Speech Recognition," *ETRI J.*, vol. 36, no. 3, June 2014, pp. 514–517.
- [3] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A Voice Activity Detection Based on the Adaptive Integration of Multiple Speech Features and a Signal Decision Scheme," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar. 31–Apr. 4, 2008, pp. 4441–4444.
- [4] J. Sohn, N.S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, Jan. 1999, pp. 1–3.
- [5] J. Ramirez et al., "Statistical Voice Activity Detection Using a Multiple Observation Likelihood Ratio Test," *IEEE Signal Process. Lett.*, vol. 12, no. 10, Oct. 2005, pp. 689–692.
- [6] T. Hughes and K. Mierle, "Recurrent Neural Networks for Voice Activity Detection," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, May 26–31, 2013, pp. 7378–7382.
- [7] Q.H. Joe et al., "Statistical Model-Based Voice Activity Detection Using Support Vector Machine," *IET Signal Process.*, vol. 3, no. 3, May 2009, pp. 205–210.
- [8] D. Enqing et al., "Applying Support Vector Machines to Voice Activity Detection," *IEEE Int. Conf. Signal Process.*, Beijing, China, vol. 2, Aug. 26–30, 2002, pp. 1124–1127.
- [9] C.Y. Park et al., "Integration of Sporadic Noise Model in POMDP-Based Voice Activity Detection," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 14–19, 2010, pp. 4486–4489.
- [10] H. Chung, S.J. Lee, and Y.K. Lee, "Endpoint Detection Using Weighted Finite State Transducer," *Proc. INTERSPEECH*, Lyon, France, Sept. 25–29, 2013, pp. 700–703.
- [11] M. Mohri, F. Pereira, and M. Riley, "Weighted Automata in Text and Speech Processing," *European Conf. AI. Intell.*, Budapest, Hungary, Aug. 13, 1996, pp. 228–231.
- [12] C. Allauzen et al., "A General and Efficient Weighted Finite-State Transducer Library," *Proc. CIAA*, Prague, Czech Republic, July 16–18, 2007, pp. 11–23.



**Hoon Chung** received his BS, MS, and PhD degrees in electronics engineering from Kangwon National University, Chuncheon, Rep. of Korea, in 1994, 1996, and 2007 respectively. He joined the Electronics and Telecommunication Research Institute, Daejeon, Rep. of Korea, in 2004 and is currently a research member of their Automatic Speech Translation and Artificial Intelligence Research Center. His current research interests include fast decoding, robust speech recognition, and large vocabulary speech-recognition systems.



**Sung Joo Lee** received his BS and MS degrees in electronic engineering from Pusan National University, Rep. of Korea, in 1996 and 1998, respectively. After graduation, he joined Hyundai Electronics Multi-media Research Center, Incheon, Rep. of Korea. Since 2000, he has been with the Electronics and Telecommunication Research Institute, Daejeon, Rep. of Korea and is a principle researcher at their Automatic Speech Translation and Artificial Intelligence Research Center. His research interests include environment-robust speech signal processing and speech recognition.



**Yun Keun Lee** received his BS and MS degrees in electronic engineering from Seoul National University, Rep. of Korea, in 1986, and Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 1988, respectively. He received his PhD in information and communication engineering from KAIST, Seoul, Rep. of Korea, in 1998. Currently, he is in charge of the Automatic Speech Translation and Artificial Intelligence Research Center at the Electronics and Telecommunication Research Institute, Daejeon, Rep. of Korea. His research interests include speech recognition, speech synthesis, and speech enhancement.