Enhanced Voice Activity Detection Using Acoustic Event Detection and Classification

Namgook Cho, *Member*, IEEE and Eun-Kyoung Kim

Abstract — *We examine user-friendly voice interface that requires the hands-free speech acquisition in the continuously listening environment. The traditional voice activity detection (VAD) algorithms cannot successfully identify potential acoustic event sounds from speech. This makes the speech recognition system frequently or incorrectly activated. In this paper, we propose a novel voice activity detection technique that consists of two major modules: 1) classification and 2) detection module. In the classification module, we label the successive audio segments based on the training models. Then, in the detection module, we remove the acoustic event sounds and make decision of the explicit utterance boundary from the input audio stream. As a result, the proposed technique enables the efficient operation of speech recognition in the continuously listening environment without any touch and/or key input. Experiments in a real-world environment and performance comparison with state-of-the-art techniques are conducted to demonstrate the effectiveness of the proposed technique*[1].

Index Terms — Acoustic event detection and classification, voice activity detection, voice interface, continuously listening environment

I. Introduction

To activate speech signal capture for speech recognition, the push-to-talk has been widely used in handheld mobile devices in which one can push a special button in the device to activate or deactivate speech capture [1,2]. It is immune to the environmental noise, but may not the case practically required in other consumer devices; for example, an interactive digital TV is supposed to listen all the time and automatically detect *only* human voice by means of the hands-free speech acquisition. In addition, it requires filtering out potential acoustic event sounds (e.g., hand-claps, phone ringing, door-slam, and so on) for the robustness of speech recognition system.

For these requirements of consumer devices, various types of voice activity detection (VAD) algorithms have been proposed to discriminate speech/non-speech sounds. Earlier algorithms for VAD are mostly based on energy levels [3], zero crossing rate [4], the periodicity measure [5], and higher order statistics in the linear prediction coding [6]. More recently, the statistical model-based VAD algorithm [7,8] has been presented, which has the decision rule derived from the likelihood ratio test. Most of these VAD algorithms assume that the background noise statistics are stationary over a longer period of time than those of speech. However, salient features of the acoustic event sounds deviate from the assumption. Thus, the VAD algorithms cannot successfully identify the acoustic event sounds from speech, which makes the speech recognition system *frequently* or *incorrectly* activated.

To address these problems, we adopt acoustic event detection and classification (AED/C) [9] and propose enhanced-voice activity detection (E-VAD) that can provide the following capabilities; 1) decision of the explicit utterance boundary from the input audio stream, and 2) removal of the potential acoustic event sounds in the continuously listening environment. E-VAD consists of two major modules; namely, classification and detection. The classification module labels the successive audio segments based on the training models. Then, the detection module removes the acoustic event sounds and determines explicitly speech boundaries for the speech recognition system. To present the effectiveness of the proposed technique, we develop a distant-talking speech recognition system operating in real time, which can control consumer devices using human voices. E-VAD enables the efficient operation of speech recognition by removing unnecessary processing of input audio stream.

The rest of this paper is organized as follows. Related previous work on AED/C is reviewed in Sec. II. The proposed E-VAD technique is described in Sec. III, where two major modules, i.e., classification and detection modules, are detailed. Experimental results are shown in Sec. IV to illustrate the superior performance of the proposed technique comparing to the conventional VAD algorithms. Finally, concluding remarks are given in Sec. V.

II. Acoustic Event Detection and Classification

The human activity is reflected in a rich variety of acoustic events, produced either by the human or by objects handled by humans, so the determination of both the identity of sounds and their position in time can help detect and describe that human activity [9]. A recent discipline, AED/C has proposed the methodology of detection and classification of acoustic event sounds in real seminar conditions. The pioneering work includes the following characteristics.

- Contrary to the previous methods for classification of sounds with a limited number of classes, e.g., speech, music, and environment sounds, AED/C handles a large and variety of acoustic event classes.

[1] Part of this work has been accepted for presentation at ICCE 2011 in Las Vegas, USA. The authors are with the Digital Media and Communications Research and Development Center, Samsung Electronics, Suwon 443-742, Korea (e-mails: namgook.cho@samsung.com and ekkim.kim@samsung.com).

Manuscript received 01/02/11
Current version published 02/21/11
Electronic version published 02/21/11.

- Since the nature of acoustic events is different from speech and music, AED/C examines the choice of appropriate features.
- Far-field microphones are used to record the acoustic event sounds in real meeting rooms.
- For the AEC and AED tasks, four evaluation campaigns were organized and the results were presented; namely, the dry-run evaluation in 2004 [9], the CHIL (Computers in the Human Interaction Loop) international evaluation in 2005 [9], CLEAR (Classification of Events, Activities, and Relationships) 2006 [10], and CLEAR 2007 [11].

A. Acoustic Event Classification (AEC)

The classification deals with acoustic events that have been already extracted or, alternatively, the temporal positions of acoustic events in an audio stream are assumed to be known. Systems proposed in the evaluation campaigns [10] can be roughly classified into three categories: GMM-, HMM-, and SVM-based systems. The SVM-based system proposed by The Technical University of Catalonia obtained the best error rate where the system exploited as segment-level features statistical parameters that are calculated from perceptual features and frequency-filtered bank energies.

B. Acoustic Event Detection (AED)

The task of AED aims to identify both time-stamps and types of acoustic events in an audio stream. Thus, it is known that AED is much more challenging than AEC. The approaches for AED can be categorized into two different ways; namely, detection-by-classification and detection-and-classification [12]. More specifically, the detection-by-classification performs detection of acoustic events by classifying the consecutive audio segments. On the other hand, the detection-and-classification finds endpoints of sounds and then performs classification with the end-pointed audio segments.

In the CLEAR 2007 evaluation [11], HMM-based system proposed by University of Illinois at Urbana-Champaign performed best among five participants, with only about 36.3% of accuracy. On average, more than 71% of all error time occurred in overlapped segments where acoustic events co-existed with speech or other acoustic events. If the temporal overlaps were not scored, the accuracy of the system would be around 60-70%. Those results confirm the difficulties of AED task in the spontaneous environment of meeting rooms.

III. Enhanced Voice Activity Detection

In this paper, we will focus on consumer devices which are supposed to work in the continuously listening environment, and to perform hands-free speech acquisition and removal of acoustic event sounds. To this end, we adopt the AED/C techniques described in Sec. II, and propose the enhanced voice activity detection.

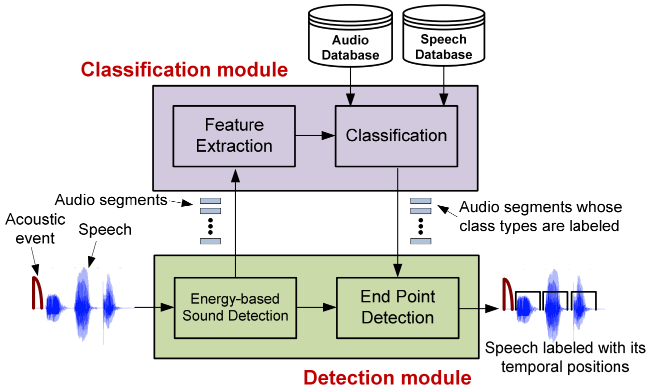

Fig. 1. The overall block-diagram of the enhanced voice activity detection that can remove acoustic events from the continuous audio stream.

E-VAD listens all the time and automatically detects the presence or absence of human speech and acoustic events. It aims to filter out acoustic events and silence, but to yield the ultimate decision on the utterance boundary for the speech recognition system. As depicted in Fig. 1, the proposed scheme consists of two major modules: classification and detection modules, which are detailed below.

A. Energy-based Sound Detection

An audio stream that may have acoustic events and/or speech is first fed to the energy-based sound detection, as shown in Fig. 1. It is then divided into non-overlapping small segments, called audio segments, using rectangular moving window of unit height; at the sampling frequency of 8000 Hz, the segments are of 1040 samples (130 msec) each.

The energy level of each audio segment is calculated as

$$e_1 = \sqrt{\sum_{n=0}^{N-1} x(n)^2}$$

where N is the size of the audio segment. An audio segment is sent to the feature extraction for classification if $e_1 \geq \eta_1$ where η_1 is a pre-defined threshold; otherwise, we drop the audio segment and move onto the next one. Therefore, the energy-based sound detection can eliminate unnecessary use of processing resources.

B. Feature Extraction

Each audio segment delivered from the energy-based sound detection is further divided into overlapping frames; at the sampling frequency of 8000 Hz, the frames are of 256 samples (32 msec) each, with 50% overlap in each of the two adjacent frames. These frames are then Hamming-windowed. A frame is marked as a silent frame if $e_2(m) < \eta_2$ where $e_2(m)$ is the normalized root mean square level of a Hamming-windowed frame, m is a frame index, and η_2 is an empirical threshold. Silent frames are not considered for classification and automatically removed.

For the discriminant features that can capture the inherent characteristics of an audio signal, we employ two types of features: perceptual features and Mel-Frequency Cepstral Coefficients (MFCCs). The perceptual features include spectral centroid, spectral flatness, spectral flux, spectral roll-off, and zero crossing rate [13,14].

- Spectral centroid (SC) measures the brightness of a sound. The higher the centroid, the brighter the sound.
- Spectral flatness (SFN) indicates how flat the spectrum of a sound is.
- Spectral flux (SF), also known as the delta spectrum magnitude, measures local spectral change.
- Spectral roll-off (SRF) quantifies the frequency value at which the accumulative value of the frequency magnitude reaches a certain percentage of the total magnitude.
- Zero crossing rate (ZCR) measures the amount of high-frequency energy.
- The first 24 MFCCs and their first derivatives are calculated for the salient features (we include log energy in the calculation).

Each analysis frame for the feature extraction is thus represented by a 55-dimensional vector.

C. Classification

1) Audio and Speech Database

To consider various types of acoustic events, we built large-scale acoustic event models using the RWCP Sound Scene Database in Real Acoustic Environment [15] and real recordings. The RWCP-DB, recorded in anechoic room, includes types of 105 acoustic events and approximately 100 samples for each type. It is grouped into three main categories and 14 sub-categories. For the real data acquisition, five types of acoustic event sounds were recorded in three different size of meeting rooms; hand-claps, cough, book-dropping, door-slam, desk. Following data collection, we categorized all the data into 17 audio classes to build the non-speech models. To build speech model, we used ETRI Korean Corpus of 1.8 million words [16]. For more details, see Sec. IV-A.

2) Classifiers

We examine two statistical learning algorithms called Gaussian Mixture Models (GMMs) and Support Vector Machine (SVM). Training for classification models the two main classes; namely, the speech class and the audio class (or non-speech class) which represents the environment acoustic characteristics. The non-speech class includes different types of acoustic event classes such as clap, door-slam, and so on. Testing labels the successive audio segments based on the training models.

As for GMM, the parameters of the models are estimated using the traditional Expectation-Maximization algorithm. Classification is made using the Maximum A Posteriori (MAP); the mean *a posteriori* log-probability on an audio segment is computed for each class model. The audio segment is then labeled according to speech or acoustic event that has the maximum *a posteriori* score.

For SVM, we use the one-against-one algorithm to construct the multi-class classifier, which is more suitable for practical use [17]; namely, $K(K-1)/2$ classifiers are built where K is the number of classes and each one is trained on data from two classes. We use kernel functions to construct the maximum-margin hyperplanes in the feature space and then perform cross-validation for finding the best parameters of the kernel function and for avoiding the overfitting problem. For testing data, we adopt the max-wins voting strategy [18,17]; every classifier assigns the data to one of the two classes, then the vote for the assigned class is increased by one vote, and finally we predict the data is in the class with the largest votes.

For temporal modeling of an audio stream, we compute the statistical features, i.e., the mean, standard deviation, entropy and auto-correlation of the feature vectors obtained in Sec. III-B. In testing, thus, we compute one statistical feature vector on an audio segment and predict the audio segment is in the class with the largest vote.

3) Confidence Measure

For classification, GMM that yields the highest log likelihood is selected as the output class. The likelihoods from the two highest scoring GMMs can be used to form a confidence measure. Let L_1 be the log likelihood of the best matching model, and L_2 be the log likelihood of the second best matching model. The confidence measure for the classification can then be computed as

$$\text{conf} = \left| \frac{L_1 - L_2}{L_1} \right| .$$

Similarly, we can apply the confidence measure to SVM in which similarity (or accuracy) scores of the two highest values are used.

By selecting an appropriate threshold, the confidence measure computed after classification can be used to modify the output of the classification. If the confidence is below the threshold, the result can be labeled as uncertain.

D. End Point Detection

Based on the classification results, the segmentation of an audio stream is achieved. Post-processing scheme is then applied to further reduce misclassification. The detailed processing flow is described below.

1) Real-time Processing

To achieve real-time processing with continuous audio streams, five signal buffers and the corresponding index buffers are prepared. Each signal buffer can save one audio segment and its classification result is marked on the corresponding index buffer. Given a new audio segment, all the audio signals in the signal buffers and their classification results in the index buffers (except the last signal and index buffers) are moved onto the next ones.

2) Speech/Non-speech Discrimination

Since each audio segment is classified into one of 16 audio classes and speech class in the classification module, the first step for explicit end point detection is to discriminate speech and non-speech segments; specifically, audio segments classified into one of 16 audio classes is labeled as non-speech segments. The discrimination results are marked in the index buffers. Note that silent audio segments and/or uncertain audio segments, which are determined in Sec. III-A and Sec. III-C, respectively, are classified as non-speech segments.

3) Smoothing

Considering that the audio stream is always continuous, it is highly impossible to change the audio class too suddenly or too frequently. Under the assumption, we apply smoothing in the final labeling of an audio sequence as

$$\text{Rule: if } s[1] \neq s[2] \ \& \ s[3] = s[1], \text{ then } s[2] = s[1],$$

where three consecutive audio segments are considered, $s[0]$, $s[1]$, $s[2]$ stands for the index of previous two and current audio segments, respectively. The rule implies that if the middle index is different from the other two while the other two are the same, the middle one is considered as misclassification. For example, if we detect a pattern of consecutive sequence like "speech– non-speech– speech," it is most likely the sequence should belong to all speeches.

4) Beginning & Hang-over

When enough segments are classified as speech in the index buffer, utterance's beginning is notified. The end point detection then includes enough segments before the beginning segment to minimize the possible detection error. Similarly, when enough non-speech segments are detected, we keep enough segments for the speech recognition before declaring hang-over, i.e., the end of utterance has been reached.

Finally, we determine explicitly utterance boundary and send only the speech segments to the speech recognizer; namely, we copy the last signal buffer to the final output buffer if its index is labeled as speech. Otherwise, we push a zero-buffer with the same size of the segment.

IV. Experimental Results

In this section, we first describe the database used to evaluate the proposed E-VAD technique. Then, we provide implementation details as well as the algorithm parameters. In addition, the performance comparison with state-of-the-art techniques is conducted. Finally, we present the application of the E-VAD technique to the distant-talking speech recognition system that operates in real time.

A. Database Description for Training

As described in Sec. III-C, we used RWCP Sound Scene Database and ETRI Korean Corpus for clean database. To consider a real-world environment (e.g., room reverberation) within which the distant-talking speech recognition system would be operated, a portable microphone/recorder was used to capture the real audio data. The recordings were made in different sizes of meeting rooms while the database sounds were played using speaker(s). We used both the recordings of real data and the clean audio clips obtained in the RWCP-DB and ETRI-DB to construct a representative, large and sufficiently diverse database. The audio clips are of different length, ranging from one second to several minutes. The selected sound classes are given in Table I. The number of items in each class is deliberately not equal, and sometimes very different. All signals in the database were downsampled to 8000 Hz, mono-channel and 16-bits per sample.

TABLE I
Classes of Sounds and Number of Samples in the Database Used for Performance Evaluation

Class	Class number	Total number	Duration (secs)
Wood	C1	1290	1037.03
Metal	C2	1000	651.82
Plastic	C3	550	413.70
Ceramic	C4	800	552.13
Dropping	C5	306	938.70
Jetting	C6	200	222.78
Rubbing	C7	500	268.88
Bursting	C8	200	101.24
Clapping	C9	939	1009.45
Small-metal	C10	1072	1294.13
Paper	C11	400	481.33
Instrument	C12	1079	1099.23
Electronic	C13	705	897.49
Mechanical	C14	1000	846.26
Cough	C15	92	472.63
Door-slam	C16	99	435.82
Speech	C17	112	1406.13
Total		10344	12128.76

B. Audio Classification Evaluation

In this subsection, we present implementation details of the classification module in Fig. 1 and examine its classification performance. In the classification evaluation, the audio clips in each class were partitioned into four sets for training and one set for testing. Each set was randomly selected. We performed a five-fold cross validation to determine parameters of the classification algorithms.

As described in Sec. III-B, a concatenation of the perceptual features and MFCCs was extracted from an audio segment to form 55-dimensional feature vectors. As for SVM, the four kinds of statistical features were computed using the 55-dimensional feature vectors. Then, we concatenated the statistical features to build one feature vector per one audio segment. In our experiments, we set $\eta_2 = 0.1$ to remove silent frames after the feature extraction.

With GMMs, each class was modeled as a mixture of several Gaussians. To determine the model order of GMM, we examined the classification results by varying the number of mixtures. Using the same settings as the rest of the experiments, we examined mixtures of 5– 30 and used the same number of mixtures for each audio class. The overall classification rates for 16 audio classes (C1– C16), are given in Fig. 2. We see that the classification performance slowly saturates as the number of mixtures increases. Due to the implementation for real-time operation, 17 mixtures were selected to model each class and each Gaussian was represented by the mean and diagonal covariance matrix of audio data. It should be noted that the appropriate number of Gaussians for each class can be estimated by the Bayesian Information Criterion (BIC) [19] as well.

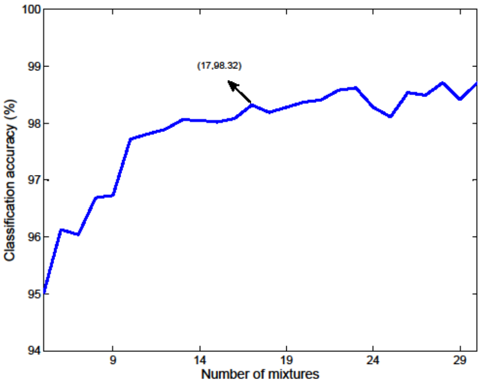

Fig. 2. Classification accuracy using GMM with a varying number of mixtures using the perceptual and MFCC features.

We also investigated the SVM classification technique. To avoid difficulties in numeric ranges, we performed linearly scaling each data to the range $\left[-1,+1\right]$. Then, we used the radial basis function (RBF) kernel to map nonlinearly data into a higher dimensional feature space. Next, we performed υ -fold cross-validation to find the best parameter of the kernel function and the best penalty parameter of SVM [17]. Finally, the best parameters were used to train the whole training set.

Table II shows the performance of the classification module of E-VAD on 10 different types of acoustic event sounds (C1, C2, C4, C5, C9– C14), where we selected *only* clean data from RWCP-DB. The classification performance was accessed in terms of average accuracy rate defined as the number of correctly classified acoustic events divided by the total number of acoustic events. We see that both GMM and SVM perform well with the clean data.

TABLE II
CLASSIFICATION RATE FOR CLEAN DATA OBTAINED FROM RWCP-DB

GMM	SVM
98.98 %	99.05 %

Next, we consider the effect of the real-room environment and training with a mixture of clean and real data on the classification performance. For the experiments, five different types of acoustic event sounds were selected; namely, C1, C3, C4, C9, and C10. First, we used only clean data for the training/testing to compare the classification performance of GMM and SVM. Then, given the models built from the clean data, the real data recorded in meeting rooms were used for testing. With the mismatched models, the classification performance degraded significantly in both GMM and SVM. To compensate the mismatched condition, we trained the classifiers with a mixture of clean and real data. The results are given in Table III. Note that SVM performs slightly better than GMM in the case of training and testing with Mixed and Real acoustic event data, respectively.

TABLE III
CLASSIFICATION RATE FOR THE EFFECT OF REAL ROOM ENVIRONMENT AND TRAINING WITH CLEAN OR MIXED DATA

Training / Testing data	GMM	SVM
Clean / Clean data	99.12 %	99.32 %
Clean / Real data	76.86 %	76.37 %
Mixed / Real data	94.14 %	97.46 %

C. Speech/Non-speech Detection Evaluation

We evaluate the performance of E-VAD on determining speech/non-speech periods from the input audio stream. In our experiments, we set $\eta_1 = 0.15$ for the energy-based sound detection.

For the discrimination analysis with real audio data, we made recordings, as shown in Fig. 3 (a), using speech utterances and eight different types of acoustic event sounds (C1, C2, C4, C5, C9, C10, C12, and C13). Then, we manually labeled the whole recordings as speech or non-speech audio regions for reference. The test signals recorded were categorized into three types; namely, signals that consist of acoustic event (AE) and speech, signals that contain acoustic event (AE) only, and signals that include speech only. We assume that there are no two or more audio types overlapped each other in the audio stream. The performance is measured as the percentage number of speech frames that are correctly recognized as speech frames.

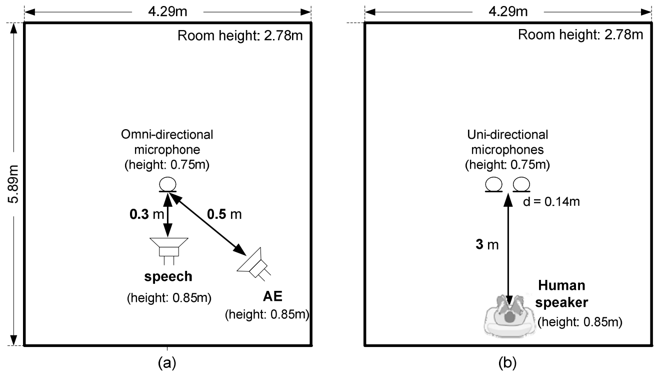

Fig. 3. The configuration of loud speakers and microphone(s) in a room; Evaluation for (a) detection and (b) the distant-talking speech recognition system. The microphones are located at 2(W)X 1.75(L)X 0.75(H)(m). AE stands for acoustic event.

Table IV presents the performance of the speech/non-speech discrimination of E-VAD. We see that GMM performs better than SVM. In the experiments, we observed that SVM produced usually discrimination errors at utterance ending.

TABLE IV
DETECTION RATE FOR SPEECH AND NON-SPEECH DISCRIMINATION

GMM	SVM
94.91 %	89.89 %

In Table V, we compare the detection performance of E-VAD with that of the conventional VAD method that adopts energy level detection [3]. Since it employs simple energy level detection, the VAD algorithm fails when the acoustic events are included in the input audio stream. The proposed E-VAD technique successfully removed acoustic events and extracted speech from the audio stream, thus it achieved high discrimination accuracy. Note that GMM was used for classification in the experiments.

TABLE V
COMPARISON FOR SPEECH/NON-SPEECH DISCRIMINATION BETWEEN E-VAD AND THE CONVENTIONAL VAD ALGORITHM

Test case	E-VAD	VAD [3]
AE + Speech	94.91 %	45.11 %
AE only	94.50 %	50.13 %
Speech only	99.80 %	93.44 %

D. Application to Distant-Talking Speech Recognition

In this subsection, we present the application of E-VAD to the distant-talking speech recognition system that operates in real time. It allows us to issue voice commands to control consumer devices such as an interactive digital TV [20]. Thus, it is supposed to listen all the time and automatically detect only human voice. E-VAD allows the system to remove from continuous audio stream acoustic event sounds which might be exist in real environments. Fig. 3 (b) illustrates the room configuration in which human speaks in front of two microphones (we assume that they are built in an interactive digital TV) while acoustic events can occur in any place in the room.

To evaluate the performance of the system, we constructed two GMM models; namely, one modeled using five acoustic classes and speech (C1, C5, C9, and C15– C17), and the other using all classes (C1– C17). Then, in each experiment session, two participants performed experiments together. More specifically, one spoke one command word(s) from a set of 26 pre-defined commands, while the other made one of the five acoustic events. Five sessions in total were performed by five pairs of different people. Table VI shows the results which were averaged over 780 trials (i.e., 26 command words × 6 test cases × 5 experiment sessions). The recognition performance was accessed in terms of average accuracy rate defined as the number of correctly recognized commands divided by the total number of command words. We see that the large and sufficiently diverse acoustic event modeling of GMM17 is able to complement the detection error of E-VAD and eventually yields the best overall performance.

TABLE VI
RECOGNITION RATE FOR 26 COMMAND WORDS USING GMM MODELED WITH SIX AND 17 SOUND CLASSES

Test case	GMM6	GMM17
Clap + Speech	91.35 %	97.44 %
Cough + Speech	85.58 %	89.74 %
Book-dropping + Speech	90.38 %	93.59 %
Door-slam + Speech	85.58 %	98.72 %
Desk + Speech	86.54 %	82.05 %
Speech only	95.19 %	96.15 %
Total	89.10 %	92.95 %

V. CONCLUSION

We examined technical barriers in the user-friendly voice interface to consumer electronics working in the continuously listening environment. Since the conventional VAD algorithms could not successfully identify the potential acoustic event sounds from speech, we adopted the AED/C techniques and developed E-VAD that can remove a variety of acoustic events and determine explicitly speech boundaries from continuous audio streams. The experimental results and performance comparison with the conventional VAD algorithm showed efficient operation of speech recognition and less recognition errors in the continuously listening environment.

REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.

[2] H. Lee, S. Chang, D. Yook, and Y. Kim, "Voice trigger system using keyword and speaker recognition for mobile devices," *IEEE Trans. Consum. Electron.*, vol. 55, no. 4, pp. 2377-2384, Nov. 2009.

[3] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Lett.*, vol. 36, no.2, pp. 180-181, 2000.

[4] J. C. Junqua, B. Reaves, and B. Mark, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Eurospeech*, 1991, pp. 1371-1374.

[5] R. Tucker, "Voice activity detection using a periodicity measure," *Proc. Inst. Electr. Eng.*, vol. 139, pp. 377-380, Aug. 1992.

[6] E. Nemer, R.Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 217-231, Mar. 2001.

[7] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1-3, Jan. 1999.

[8] J. Chang, N. K. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no.6, pp. 1965-1976, Jun. 2006.

[9] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, Acoustic Event Detection and Classification, *Computers in the Human Interaction Loop*, A. Waibel and R. Stieflhagen (Eds.), Springer, 2009.

[10] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "CLEAR evaluation of acoustic event detection and classification systems," *CLEAR'06 Evaluation Campaign and Workshop*, Southampton, UK, 2006, in Multimodal Technologies for Perception of Humans, LNCS, vol. 4122, pp. 311-322, Springer, 2007.

[11] A. Temko, C. Nadeu, and J-I. Biel, “Acoustic event detection: SVM-based system and evaluation setup in CLEAR’07”, *CLEAR’07 Evaluation Campaign and Workshop*, Baltimore, MD, USA, 2007, in Multimodal Technologies for Perception of Humans, LNCS, vol. 4625, pp. 354-363, Springer, 2008.

[12] A. Temko, “Acoustic Event Detection and Classification,” Ph.D. dissertation, Technical University of Catalonia, Barcelona, Spain, 2008.

[13] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition with time-frequency audio features,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no.6, pp. 1142-1158, Aug. 2009.

[14] K.-M. Kim, S.-Y. Kim, J.-K. Jeon, and K.-S. Park, “Quick audio retrieval using multiple feature vectors,” *IEEE Trans. Consum. Electron.*, vol. 52, no. 1, pp. 200-205, Feb. 2006.

[15] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano, “Environment sound source identification based on hidden Markov model for robust speech recognition,” *Proc. EuroSpeech*, pp. 2157-2160, 2003.

[16] Electronics and Telecommunications Research Institute (ETRI) Korean Corpus, 2005.

[17] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415-425, Mar. 2002.

[18] J. Friedman, “Another approach to polychotomous classification,” Tech. Report, Dept. Statist., Stanford Univ., 1996.

[19] C. Fraley and A. E. Raftery, “How many clusters? Which clustering method? Answers via model based cluster analysis,” Tech. Report, Dept. Statist., Univ. of Washington, 1998.

[20] X. Zeng, A. D. Fapojuwo, and R. J. Davies, “Design and performance evaluation of voice activated wireless home device,” *IEEE Trans. Consum. Electron.*, vol. 52, no. 3, pp. 983-989, Aug. 2006.

BIOGRAPHIES

Namgook Cho (S’07-M’10) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, in 2009. Since 2009, he has been with the Digital Media and Communications R&D Center, Samsung Electronics, Suwon, Korea. His research interests include sparse representations of signals, audio source separation, and signal processing and machine learning for content analysis.

Eun-Kyoung Kim received the M.S and Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, Republic of Korea, in 1997 and 2003. Since 2003, she has been with the Digital Media and Communications R&D Center, Samsung Electronics, and worked to develop speech interface for mobile & CE devices. Her research interests include robust speech recognition, statistical speech synthesis, and mobile voice applications.