Embedded Discriminant Analysis based Speech Activity Detection for Unsupervised Stress Speech Clustering

Barlian Henryranu Prasetio
Interdisciplinary Graduate School of Agriculture and Engineering
University of Miyazaki
Miyazaki, Japan
Email: barlian@ub.ac.id

Hiroki Tamura
Faculty of Engineering
University of Miyazaki
Miyazaki, Japan
Email: htamura@cc.miyazaki-u.ac.jp

Koichi Tanno
Faculty of Engineering
University of Miyazaki
Miyazaki, Japan
Email: tanno@cc.miyazaki-u.ac.jp

***Abstract*—Speech activity detection (SAD) or sometimes called voice activity detection (VAD), is a crucial part of most speech-related applications. The SAD system serves to ensure the primary system processes only speech segments. Many speech-based systems have reported that detection accuracy is thanks to the robustness of their SAD system. Various SAD methods have been explored and enhanced in addressing noisy environments, but a few of them notice the emotional condition of the speakers. Whereas, in real conditions, emotions (such as stress) can pose a considerable impact on SAD system performance. In this paper, we propose a compact SAD system that is able to harmonize with the altered speech characteristics due to the presence of emotion and also powerful in high noise conditions. Since there is a similarity between emotional effect and channel effect, the advantages of the proposed SAD system is the applied of a new channel compensation scheme (termed as embedded discriminant analysis, EDA) that works in the i-vector space. We design the EDA in such a way so that it could compensate the presence of emotional condition. EDA transforms original i-vector to a lower-dimensional denoise embedding space. We develop EDA as simple and efficient as the linear discriminant analysis (LDA). The cosine similarity algorithm is applied to calculate the resemblance score between the audio target and the speech/non-speech models, and also for deciding the decision threshold. The effectiveness of the proposed SAD system is evaluated in the clustering task of Speech Under Simulated and Actual Stress (SUSAS) data, that aimed for the stress speech clustering (SSC) system.**

***Contribution*—We propose a SAD system which not only strong in noisy environments but also be able to compensate the presence of emotional conditions.**

***Index Terms*—Embedded discriminant analysis, SAD system, Stress conditions, I-vector, Neural network.**

I. Introduction

Speech activity detection (SAD) is an essential part of the speech-based applications [1]. SAD plays a critical role in separating speech dan non-speech segments, such as noise, music background, or silence. Typically, SAD is performed as the first task to filter the presence of non-speech segments. Non-speech segments considerably affect the performance of the main system due to carry useless information. In other words, a robust SAD system is an important modal to obtain an accurate detection result of the main system.

Many feature extraction techniques that can reflect the speech features. For instance, energy-based features [2], [3] and Mel-frequency cepstral coefficients (MFCCs) [4]–[7] are the technique that is frequently used in representing the presence of speech. Both techniques are robust in clean conditions, but the performance degrades in noisy environments. In this decade, the learning-based technique has been explored in the term of extracting the speech features [7]. The GMM-based feature and its variance have been successfully used in projecting the presence of speech [4], [6], [8]. Furthermore, an effective and sophisticated technique, known as i-vector, also has shown their ability to present a compact speech feature for a SAD system [4], [9], [10].

A linear method, such as linear discriminant analysis (LDA), has successfully discriminated against the speech and non-speech [5], [11]. More than 30-years, LDA has been commonly used as a standard back-end procedure in a wide range of speech-related tasks. By assuming each class is a Gaussian distribution and all classes share the same covariance matrix, LDA shows its effectiveness in stationary noise conditions. The machine learning model, such as the hidden Markov model (HMM) [6] and support vector machine (SVM) [4], [9], has proven more accurate in non-stationary noise conditions but involve a complex procedure.

Nowadays, deep neural networks (DNNs) have been widely used and achieves extremely high predictive accuracy in hardly overall machine learning applications, included as a feature compensation and denoising technique [12] in emotion recognition. In the denoising task, DNN learns the entire temporal context of input in-depth and learn its projection by mapping the noisy feature (i-vector) to a denoised space [13]. [3], [10], [14] explicitly use DNN as a channel compensation for a SAD system and it proven effective in a non-stationary noise.

978-1-7281-9331-1/20/$31.00

As discussed above, LDA is the most popular and preferable model due to its simpleness and efficiency in recognizing a pattern but susceptible to non-stationary noise. On the other hand, DNN shows its effectiveness as a denoise technique but involves more complex procedures. Many studies have explored various approaches to develop a SAD system. Most of them focus on how to recognize the presence of the speech in high noise conditions, and just little works notice the emotional condition of the speaker also. Whereas, the presence of emotion (especially stress condition) affects the performance of speech algorithms [15]. For a standard speech-related system, such as speech recognition [16], [17] or speaker recognition [18], [19], a robust SAD system in the noisy conditions has adequate. However, for the emotion-related system [20]–[22], a more powerful SAD system is required because the emotional condition affects the production of speech characteristics.

To this end, we intend to consider the effect of the emotional condition of the speaker in the SAD system. It was found that the emotional effect is rather similar to the channel effect [23]. Another study mentioned that it allows modeling intrinsic variability, such as emotion, using technique modeling of extrinsic variability [24]. Thus, emotion variability could be considered as a type of channel variability. It means a variability of each emotional state is predictable in the same way to each channel. Channel variability is one of the influential factors of speech based-systems successful. It has the profit that the transformed parameters can be applied to systems that have different environments [25].

In this paper, we propose a SAD system which not only strong in noisy environments but also be able to compensate for the presence of the speech that might alter because of emotional conditions. The proposed SAD system consists of the i-vector feature extractor and a novel channel compensation method, named as embedded discriminant analysis or EDA. EDA is a channel compensation method that as simple and efficient as LDA but also has an ability to transforms the feature to denoise space like the DNN. EDA distinguishes speech and non-speech effectively by mapping each frame of i-vector feature to a more discriminative space using DNN and modeling its transformation in a projection matrix like the LDA model. We explicitly use a time-delay neural network (TDNN) in the EDA's structure to handle the variations of temporal dependencies caused by the presence of emotional stress [26]. In the training phase, a large amount of short speech data from the SUSAS dataset are used to create the speech/non-speech model. In the testing phase, the cosine similarity algorithm is used to compute the deviation between the speech/non-speech model and the representation of the audio target, and also for deciding the decision threshold. Based on this threshold, the speech and non-speech boundaries are decided. The effectiveness of the proposed SAD system is evaluated in terms of the stress speech clustering task [22].

II. Related Works

A fundamental issue in the most speech-based system is how to ensure only human speech is processed to the system and ignore the other voices (non-speech). The non-speech conditions carry useless information so that it can degrade system performance. Non-speech can be interpreted as silence, noise, music background, etc. To address this issue, most of the speech-based systems eliminate the non-speech segments before the main process [1].

A classic SAD system used energy-based feature [2], [3] or spectral-based feature [3]–[6] to reflect the presence of speech. The identity vectors (i-vector) technique that commonly used for speaker recognition [7], lately also becomes sophisticate to represent the speech and non-speech. Then, to distinguish the speech and non-speech, a linear discrimination method such as LDA [5], [11] and PLDA [4] is performed. Many studies explored a machine learning approach to discriminate speech and non-speech. For instance, hidden Markov model (HMM) [6], support vector machine (SVM) [4], [9], deep neural network (DNN) [10], [14], and long short-term memory (LSTM) [3].

DNNs are proven robust in noisy conditions but drop the performance for a low signal-to-noise ratio (SNR) conditions [3], such as Speech Under Simulated and Actual Stress (SUSAS) database [27], [28]. On the other hand, another DNN type shows its effectiveness in capturing the speech information in a short and long temporal context [29], known as a time-delay neural network (TDNN). Different from DNN that processes the whole temporal context since first layers, TDNN processes a small context in the initial layer and continues in a wider temporal context on the deeper layer. Hence, we develop a TDNN-based channel compensation method for the SAD system, named embedded discriminant analysis or EDA. EDA transforms the i-vector features to a low-dimensional embedding space that more discriminative and denoised. By utilizing the TDNN structure as a backbone, EDA is able to handle the dynamic dependencies of temporal contexts in order to compensate emotion variability.

III. Proposed Method and Materials

The proposed SAD system consists of the training and testing phase, as shown in Fig. 1. In the training phase, the speech and non-speech features are extracted using the MFCC technique. Then, the i-vectors extractor transforms each frame of the speech features to a single low-dimensional i-vector space. By using the i-vector feature of speech and non-speech, EDA is trained to produce the speech and non-speech model. In the testing phase, the same procedure is conducted to the audio target. We perform the cosine similarity algorithm to compute the resemblance score between the audio target and both of speech/non-speech models. Finally, the score based-error evaluation metric is applied for deciding the decision threshold.

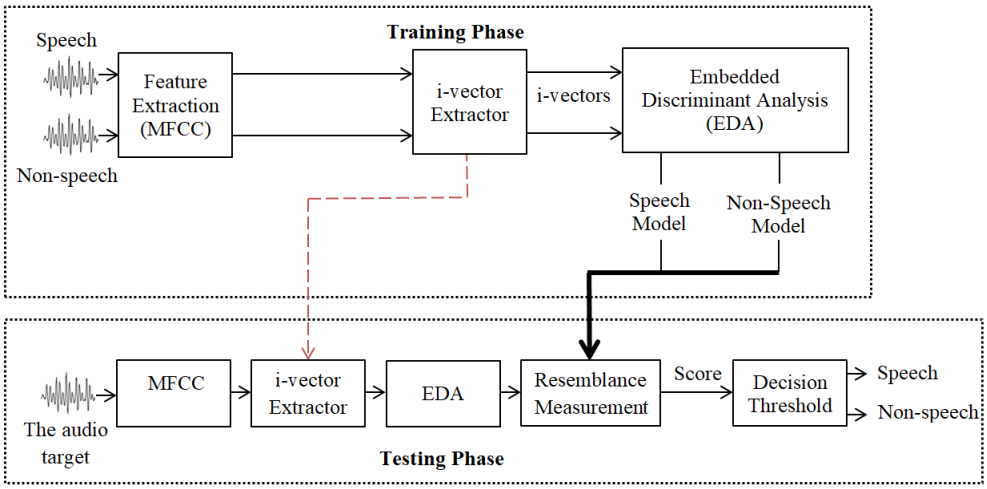

Fig. 1. The proposed end-to-end SAD system in training and testing phase

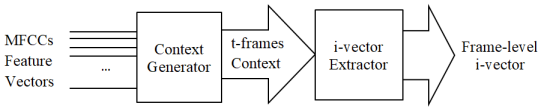

Fig. 2. The frame-level i-vector extractor

A. Feature Extraction

We use the i-vector extractor to transform as input the sequence of MFCC feature vectors to frame-level i-vector features. As shown in Fig. 2, the context generator is applied to split MFCC feature vectors to "t-frames" context. Then, on each frame, i-vector extraction is performed to produce a frame-level i-vector feature [30].

The total variability model (TVM) represents i-vector in a single low-dimensional vector model [4], which given as follows:

$$s = m + Tw \tag{1}$$

where s is a super-vector adapted to UBM, m is mean super-vector, T is the matrix representing the subspace, and w is i-vector representation.

The first order statistics ($[f_1^t, f_2^t, ...f_V^t]$, where V is the number of mixture of the UBM) of given sequence of MFCC feature vectors ($\{x_1, x_2, ...x_t\}$, where t is number of frames) are estimate to obtain i-vector representation. f_v is a subvector of f that is defined as follows [30]:

$$f_v = \Sigma_v^{-\frac{1}{2}}(\sum \gamma_{t,v} X_t - \mu_v) \tag{2}$$

where Σ_v and μ_v are the covariance and mean matrix of the v^{th} mixture of the UBM, respectively. $\gamma_{t,v}$ is the posterior of t^{th} frame of speech for v^{th} mixture. By assuming the speech vectors are normal distribution, $\gamma_{t,v}$ defined as follows [30]:

$$\gamma_{t,v} = P(X_t|\mu_v, \Sigma_x) \tag{3}$$

and for given the first order statistics, i-vector is obtained by [30]:

$$w = (I + \sum_{v=1}^{V} H_v T_v^t \Sigma_v^{-\frac{1}{2}} T_v)^{-1} T^t \Sigma^{-1} f \tag{4}$$

where $H_v = \sum_t \gamma_{t,v}$, T_v is the submatrix of T for the v^{th} mixture, and Σ is the block diagonal matrix with Σ_v as blocks along the diagonal.

B. Channel Compensation Method

1) Linear Discriminant Analysis (LDA): For more than three decades, LDA has become the most simple and effective channel compensation method. LDA transform a linear representation of a high-dimensional feature vector y into a low-dimensional discriminative subspace u that projected as W matrix, $(W : \mathbb{R}^h \rightarrow \mathbb{R}^l)$, formulated as follows:

$$u = W^T(y) \tag{5}$$

W matrix represents the inter-class S_b and intra-class S_w covariance matrix of the two classes that to be discriminated [5], [11], in our case: speech and non-speech. The LDA defines the criteria λ to separating the speech and the non-speech class, as follows:

$$\lambda = \frac{W^T S_b W}{W^T S_w W} \tag{6}$$

The projection matrix W contains the eigenvectors that correspond to the largest eigenvalue of $S_w^{-1} S_b$, which is chosen as a solution for LDA optimization.

2) Proposed Embedded Discriminant Analysis (EDA): Different to LDA that assumes all classes share the same covariance matrix and find its linear transformation using Gaussian distribution, while EDA initially maps the features to an embedding space, then finds the corresponding transformation using the trained neural network.

TABLE I
THE EDA NETWORK STRUCTURE

Layer	Input context	Dimensions	Function
Input	-	600	-
Hidden-1	[t-8,t,t+8]	600	ReLU
Hidden-2	[t-5,t,t+1]	600	ReLU+Batch-Norm
Embedding	{0}	400	
Loss layer	Jointly supervision of softmax loss and center loss		

We propose to use a one-dimensional convolutional network (known as TDNN) in the EDA structure to address the temporal dynamics dependencies caused by the emotional conditions of the speaker's [29]. TDNN learns dynamically temporal dependency by generating larger networks from sub-components at across time steps [31]. We applied the sub-sampling (locally-connected) technique on both hidden layers of EDA to make it more efficient. Structurally, the proposed EDA consists of two hidden layers, the embedding layer, and the loss layer, as shown in Table I. EDA capture a total temporal context on [-13,9] [31] that processed by 2 layers. The hidden-1 layer splice together frames $t-8$ through $t+8$ and the hidden-2 layer splices together frames $t-5$ through $t+1$. We applied a rectified linear unit (ReLU) as activation functions on all hidden layers and incorporated a batch normalization in the hidden-2 layer to stabilize the training procedure. During training, the parameters of EDA are optimized under the supervision of softmax loss and center loss.

The softmax loss function [32] is defined as follows:

$$\mathcal{L}_S = -\sum_{i=1}^{G} log \frac{e^{W_{z_i}^T y_i + b_{z_i}}}{\sum_{j=1}^{Z} e^{W_j^T y_i + b_j}} \tag{7}$$

where $y_i \in \mathbb{R}^d$ denotes i^{th} embedding feature, belonging to $z_{i^{th}}$ class, and Z denotes the number of softmax outputs (number of classes). W_j is j^{th} column of the weights matrix W and b is bias term. d and G are feature dimensions and the total number of training samples (i-vector), respectively.

The center loss function [33], defined as follows:

$$\mathcal{L}_C = \frac{1}{2} \sum_{i=1}^{G} \parallel y_i - c_{z_i} \parallel^2 \tag{8}$$

where c_{z_i} denotes the $z_{i^{th}}$ class center of embedding feature. Equation (8) shows that the intra-class variations are effectively characterized. The c_{z_i} is updated based on mini-batch and the center compute by averaging the embedding feature of the corresponding class iteratively. A scalar α is used to control the learning rate of the center where α is [0 to 1] restrictively.

After training, the transformed features are extracted from the affine component of the embedding layer, hereinafter referred to the embedding feature that is formally trained under the supervision of softmax loss and center loss, as follows:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C \tag{9}$$

where λ is a weight for the balance of two-loss functions. If λ is small, the loss is supervised by softmax loss. Otherwise, the supervision is inclined to the center loss.

EDA transforms the i-vector feature from an original space y into embedding space u. We define the transformation model of EDA in a projection function similar to the LDA model (Eq. 5), as follows:

$$u = \theta(y) \tag{10}$$

where θ denotes the transformation projection matrix of EDA.

C. Resemblance Measurement

In the training phase, the proposed SAD system is trained to generate speech and non-speech models. We stored one model vectors per trained segments. Then, those models (speech ϕ_{sp} and non-speech ϕ_{nsp}) are presented in d-dimensional vectors, defined as follows:

$$\begin{aligned} \phi_{sp} &= (\mathcal{N}_1(u_{sp}), \mathcal{N}_2(u_{sp}), ..., \mathcal{N}_d(u_{sp})) \\ \phi_{nsp} &= (\mathcal{N}_1(u_{nsp}), \mathcal{N}_2(u_{nsp}), ..., \mathcal{N}_d(u_{nsp})) \end{aligned} \tag{11}$$

where u_{sp} and u_{nsp} are speech and non-speech vectors, respectively. We then compute the deviation between the embedding feature (the output of EDA) of audio target ϕ with both models (speech ϕ_{sp} and non-speech ϕ_{nsp}) using cosine resemblance algorithm [8], formulated as follows:

$$\begin{aligned} \mathcal{S}_{sp}(\phi) &= cos(\phi, \phi_{sp}) - cos(\phi, \phi_{nsp}) \\ &= \frac{\phi^T}{\|\phi\|} \left(\frac{\phi_{sp}}{\|\phi_{sp}\|} - \frac{\phi_{nsp}}{\|\phi_{nsp}\|} \right) \end{aligned} \tag{12}$$

where $\| \, . \, \|$ denotes the euclidean distance.

D. Decision Evaluation Metrics

We evaluate the proposed system using the equal error rate (EER) [34] evaluation metrics for a two-classes classification task (speech and non-speech). This evaluation is obtained from two main errors: False Alarm Rate (FAR) and False Rejection Rate (FRR). FAR is the percentage of non-speech frames classified as speech, while FRR is the percentage of speech frames classified as non-speech, and EER is a intersect between FAR and FRR, or in other words, at which FAR and FRR are equal. We define an EER in term of the decision threshold η, as follows:

$$EER = \frac{FRR(\eta) + FRR(\eta)}{2} \tag{13}$$

IV. EXPERIMENTAL SETUP

A. Dataset

Many attempts have been made to ensure high accuracy results of speech-based systems in the framework of adaptation to noise or speaker characteristics. However, the presence of nuisance factors in high capacity and also the diversity of speech styles are still found, such as changes in speech due to the Lombard effect or speech under pressure and emotional influence. Therefore, we present an adaptation to emotional speaker conditions for a more powerful SAD system. Besides

being able to recognize the presence of the speech, the proposed SAD system shall also be proven effective on the emotional stress speech database.

In light of the challenge of the proposed SAD system effectiveness in the emotional condition, we provide a little comparative analysis of the most prevalent emotional databases. An emotional speech database is addressed for a specific purpose. Ververidis. et al. [35] reviewed 32-emotional speech databases from 13-languages. Most databases aimed at automatic speech recognition and just two-databases collected for stress recognition, i.e., SUSAS and MIT Lab [36]. Schuller. et al. [37] provide benchmark comparison on the nine corpora under equal conditions. From the overview, it shows that SUSAS provides several levels of emotional stress. Emotion databases are recorded from a natural or simulated condition. Since humans cannot easily classify natural emotions (an emotion that recorded from the natural condition), machines cannot present a higher result of classification. To avoid more arduous situations, corpora provide the simulated version of emotional speech. Two-databases offer both versions (natural and simulated), i.e., SUSAS and Scherer [38]. As discussed above, SUSAS is a qualified database for training or evaluating the speech-based system that notices the emotional condition of the speaker.

SUSAS database was constructed and introduced by [39] and it was collected by the Linguistic Data Consortium (LDC) that consisted of labeled short speech and unlabeled conversation speech [27], [28]. This database is addressed to the study in terms of speech production and recognition varies of speaking during stressed conditions. Therefore, besides recognize stress and emotion in speech [40], the SUSAS database also has been used in many speech-based systems for handling the speech spoken by under stressed speakers, such as speaker identification [41] and recognition [42], [43], automatic speech recognition [44], [45], and gender identification [46].

The effectiveness of the proposed SAD system is evaluated using short and long-duration speech data of the SUSAS database. Specifically, for the training phase, we used 1377 males and 1323 females speech data and more than 1500 non-speech data that consisted of high noise, low noise, and silence. For the evaluation phase, we used the long-duration data (conversation) that has various length duration as the audio target. We employ the annotations-based ground truth to evaluate the effectiveness of the proposed SAD system.

B. Proposed System Settings

1) I-vector Parameters: We extract the MFCC feature of the speech at a 10ms frame rate with a 25ms window size. The 13-dimensional MFCC is used as the input of the i-vector extractor to produce a frame-level i-vector feature. The UBM super-vector contains 2048 Gaussian mixtures is applied to produce a 600-dimensional i-vector.

2) EDA Parameters: As shown in Table I, EDA contains one input layer, two hidden layers, and one embedding layer. ReLU is used as an activation function and we incorporated a batch normalization layer in the second hidden layer. We set the weight balancing parameter for softmax loss and center loss $\lambda = 10^{-2}$ and the controller parameter for learning rate of center $\alpha = 10^{-1}$ [32].

TABLE II
THE PERFORMANCE COMPARISON RESULT OF THE PROPOSED SAD SYSTEM (% EER). THE PERFORMANCE IS PRESENTED FOR DIFFERENT SYSTEM AND SPEECH DURATION.

System	10-sec	30-sec	60-sec
SSC without SAD system	41.584	44.660	49.675
SSC with Baseline SAD system	32.673	33.981	36.526
SSC with Proposed SAD system	29.703	29.773	29.870

Note: SSC is stress speech clustering system [22].

C. Baseline SAD System Settings

The baseline SAD system is a SAD system (as shown in Fig. 1) which uses LDA as the discriminant model. We set the LDA's parameter setting as used in [11].

V. Result and Discussions

The effectiveness of the proposed SAD system (EDA-based SAD system) is evaluated in the task of classification of the SUSAS dataset that presented for the SSC system. We perform EDA and LDA as channel compensation method in the proposed SAD system and the baseline SAD system, respectively.

In this experiment, EDA and LDA reduce the i-vector dimension from 600 to 400. The performance comparison result of the proposed SAD system in terms of EER is shown in Table II and the example of data segmentation is presented in Fig. 3. Generally, all systems present increased error for long speech duration. The proposed SAD system outperforms baseline systems and relatively stable in all speech duration. This indicates that by using EDA, the proposed SAD system is able to capture speech information in a short and long temporal context.

Fig. 3 shows the proposed SAD system decision results compared with the baseline SAD system, with a correspondence confidence threshold score. Fig. 3(a) shows the original speech signal that has 3 types speech/non-speech conditions (silent: 0 to 0.25 seconds, speech: 0.26 to 0.95 seconds, high noise: 1.1 to 1.2 seconds, and low noise: 1.25 to 1.5 seconds). In Fig. 3(b) and (d), the SAD score is close to 0 for the silent condition, the negative score for noise conditions, and the positive score for speech conditions. Since EDA is able to handle the variations of temporal dependencies, the proposed SAD system presents more sensitive SAD scores in different non-speech conditions (Fig. 3(d)). On the other hand, due to LDA uses the Gaussian approach, some speech frames (Fig. 3(c) time: 0.25 to 0.35) are recognized as the non-speech.

Unlike the LDA-based SAD system, the proposed SAD system shows its capability in capturing the nonlinear relationship between features without requiring the prior assumptions on the input. While LDA probably less adequate due to speech under stress is not always distributed in Gaussian, and it may have different covariance. Since EDA notices emotional information on the speech, the proposed SAD system obtains

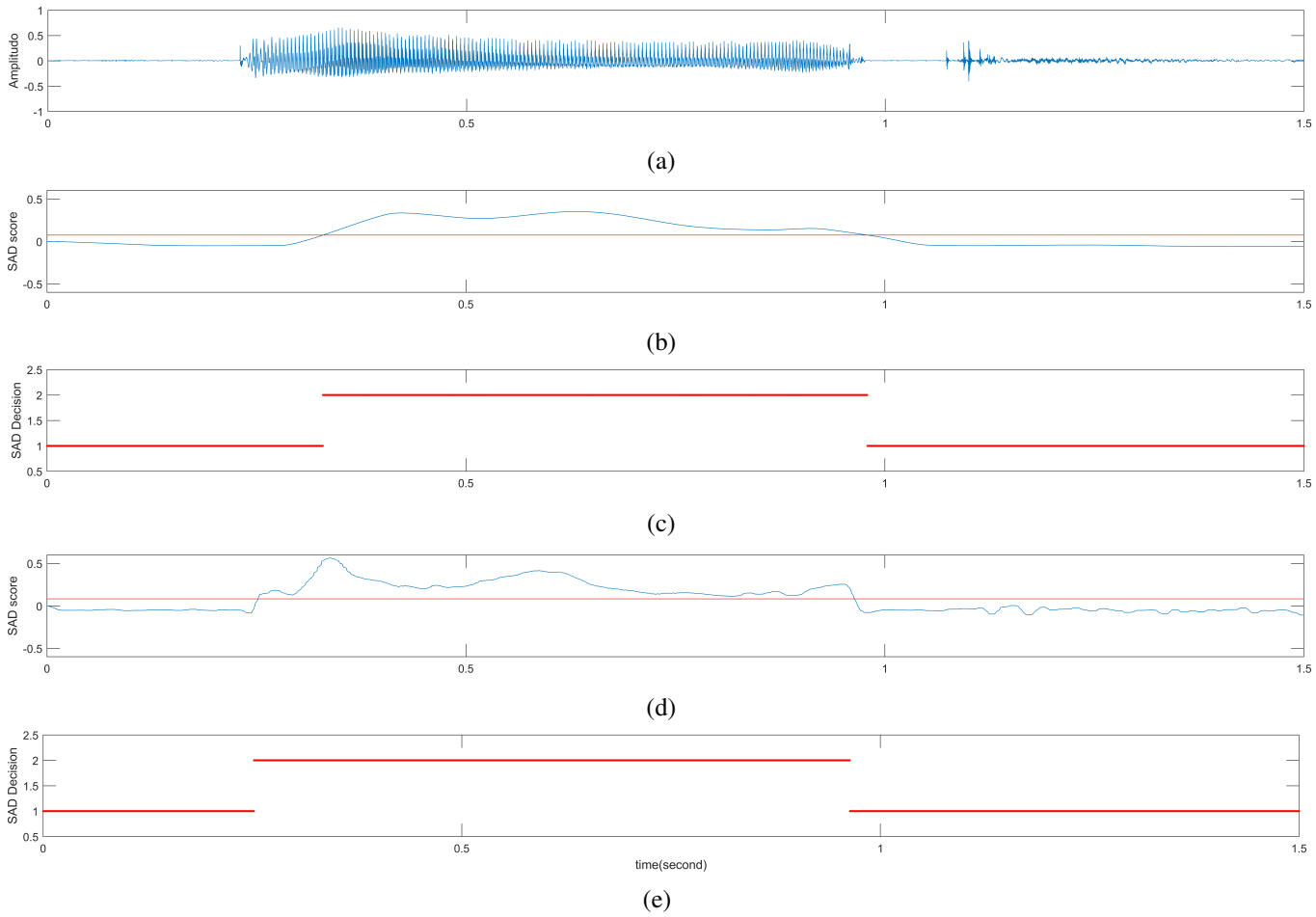

Fig. 3. The example of data segmentation in the time domain, and the corresponding SAD decision. (a) demonstrates the original signal in 1.5 seconds with frame-scale $fs = 8000$. (b) and (d) denote the SAD score for the baseline and proposed SAD system, where the blue indicates the SAD score and the red line is a correspondence threshold. (c) and (e) present the SAD decision for the baseline and proposed SAD system, respectively.

a more accurate in the SAD score that followed by an accurate decision threshold, as shown in Fig. 3. It impacts on the ability to distinguish the different non-speech types, such as silent, high noise, and low noise.

VI. Conclusion

In this paper, a compact speech activity detection (SAD) system using i-vector and proposed embedded discriminant analysis (EDA) has been presented. The proposed SAD system was not only strong in noisy environments but also be able to compensate for the presence of emotional conditions. In the training phase, the speech and non-speech features were extracted using Mel Frequency Cepstral Coefficients (MFCC) technique that was then transformed to frame-level feature by the i-vector extractor. The proposed EDA transformed the i-vector features into denoise embedding space by supervision of softmax and center loss. To compensate emotional conditions, EDA was trained using labeled short speech data of the SUSAS database to produce a projection function and was used for generating the speech/non-speech models. In the testing phase, the cosine similarity algorithm was used to computes the deviation between the speech/non-speech models and the audio target. The effectiveness of the proposed SAD system was evaluated in terms of the equal error rate (EER) by comparing it with the baseline SAD system as a pre-processing part of the stress speech clustering (SSC) system. Based on the experiment, the proposed SAD system presented a stable EER in short and long speech durations and also presented a sensitive SAD scores in different non-speech conditions. The use of a fixed threshold in all speech frames makes the speech that has different emotions get the same treatment. Therefore, the interest direction of future work is to explore the dynamic thresholding that adaptive to different emotional conditions. Thus, an algorithm that modulates the threshold adaptively would be proposed in the future work for resulting in a time-adaptive thresholding η_t, where t is time.

Acknowledgment

The authors would like to thank the LDC for allowing us access to the SUSAS database.

REFERENCES

[1] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, *Voice Activity Detection: Merging Source and Filter-based Information*, IEEE Signal Processing Letters, 23(2), 252-256, 2016.

[2] K. Sang-Kyun, K. Sang-Ick, P. Young-Jin, L. Sanghyuk, and L. Sangmin, *Power Spectral Deviation-Based Voice Activity Detection Incorporating Teager Energy for Speech Enhancement*, Symmetry, 8(58), 8 pages, 2016.

[3] P. Sertsi, S. Boonkla, V. Chunwijitra, N. Kurpukdee, and C. Wutiwiwatchai, *Robust Voice Activity Detection Based on LSTM Recurrent Neural Networks and Modulation Spectrum*, Proceedings of APSIPA Annual Summit and Conference, Kuala Lumpur, Malaysia, 2017.

[4] E. Khoury, and M. Garland, *I-Vectors for Speech Activity Detection*, The Speaker and Language Recognition Workshop (Odyssey), Bilbao, Spain, 2016.

[5] E. Rentzeperis, C. Boukis, and A. Pnevmatikakis, *Combining Finite State Machines and LDA for Voice Activity Detection*, Artificial Intelligence and Innovations (AIAI): from Theory to Applications, IFIP The International Federation for Information Processing, 247, Springer, Boston, MA, 2007.

[6] Y. Liang, X. Liu, M. Zhou, Y. Lou, and B. Shan, *A Robust Voice Activity Detector Based on Weibull and Gaussian Mixture Distribution*, International Conference on Signal Processing Systems (ICSPS), Dalian, China, 2010.

[7] P. Verma and, P. K. Das, *i-Vectors in speech processing applications: a survey*, International Journal of Speech Technology, 18(4), 529-546, 2015.

[8] O. Ghahabi, W. Zhou, and V. Fisher, *A Robust Voice Activity Detection for Real-time Automatic Speech Recognition*, The Conference on Electronic Speech Signal Processing (ESSV), Baden-Württemberg, Germany, 2018.

[9] Z. Huang, Y. Cheng, K. Li, V. Hautamaki, and C. Lee, *A Blind Segmentation Approach to Acoustic Event Detection Based on I-Vector*, INTERSPEECH, Lyon, France, 2013.

[10] H. Yamamoto, K. Okabe, and T. Koshinaka, *Robust i-vector extraction tightly coupled with voice activity detection using deep neural networks*, Asia-Pacific Signal and Information Processing Association (APSIPA), Annual Summit and Conference, Kuala Lumpur, Malaysia, 2017.

[11] J. Padrell, D. Macho and C. Nadeu, *Robust speech activity detection using LDA applied to FF parameters*, The IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA, 2005.

[12] R. Chakraborty, A. Panda, M. Pandharipande, S. Joshi, and S.K. Kopparapu, *Front-end Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition*, INTERSPEECH, Graz, Austria, 2019.

[13] R. Xia and Y. Liu, *Using Denoising Autoencoder for Emotion Recognition*, INTERSPEECH, Lyon, France, 2013.

[14] S. Dwijayanti, K. Yamamori, M. Miyoshi, *Enhancement of speech dynamics for voice activity detection using DNN*, EURASIP Journal on Audio, Speech, and Music Processing, 2018(10), 1-15, 2018.

[15] S.E. Bou-Ghazale and J.H.L. Hansen, *A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress*, IEEE Trans. on Speech and Audio Processing, 8(4), 429-442, 2000.

[16] M.F. Alghifari, T.S. Gunawan, M.A. binti Wan Nordin, S.A.A. Qadri, M. Kartiwi, and Z.Janin, *On the use of voice activity detection in speech emotion recognition*, Bulletin of Electrical Engineering and Informatics, 8(4), 1324-1332, 2019.

[17] D.Sztahó and K.Vicsi, *Speech activity detection and automatic prosodic processing unit segmentation for emotion recognition*, Intelligent Decision Technologies, 8(4), 315-324, 2014.

[18] J. Ling, S. Sun, J. Zhu, and X. Liu, *Speaker Recognition with VAD*, Pacific-Asia Conference on Web Mining and Web-based Application, Wuhan, China, 2009.

[19] M. Mak and H. Yu, *A study of voice activity detection techniques for NIST speaker recognition evaluations*, Computer Speech & Language 28(1), 295–313, 2014.

[20] B.H. Prasetio, H. Tamura, and K. Tanno, *Ensemble Support Vector Machine and Neural Network Method for Speech Stress Recognition*, International Workshop on Big Data and Information Security (IWBIS), Jakarta, Indonesia, 57-62, 2018.

[21] B.H. Prasetio, H. Tamura, and K. Tanno, *Generalized Discriminant Methods for Improved X-Vector Back-end Based Speech Stress Recognition*, *IEEJ Transactions on Electronics, Information and Systems, 139(11*, 1341-1347, 2019.

[22] B.H. Prasetio, H. Tamura, and K. Tanno, *A Deep Time-delay Embedded Algorithm for Unsupervised Stress Speech Clustering*. Proceeding of IEEE International Conference on Systems, Man, and Cybernetics (SMC), Bari, Italy, 2019.

[23] H. Bao, M. Xu, F. Zheng, "Emotion Attribute Projection for Speaker Recognition on Emotional Speech", INTERSPEECH, pp.758-761, 2007.

[24] E. Shriberg, S. Kajarekar, and N. Scheffer, "Does Session Variability Compensation in Speaker Recognition Model Intrinsic Variation Under Mismatched Conditions?" INTERSPEECH, pp. 1551-1554, Brighton, United Kingdom, 2009.

[25] D. Colibro, C. Vair, F. Castaldo, E. Dalmasso, P. Laface, *Speaker recognition using channel factors feature compensation*, European Signal Processing Conference (EUSIPCO), Florence, Italy, 2006.

[26] O.V. Verkholyak, H. Kaya, and A.A. Karpov, *Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification*, Tr. SPIIRAN, 18(1), 30-56, 2019.

[27] J. H. L. Hansen, *Composer, SUSAS LDC99S78. Web Download. [Sound Recording]*, Philadelphia: Linguistic Data Consortium, 199).

[28] J. H. L. Hansen, *Composer, SUSAS Transcripts LDC99T33. [Sound Recording]*, Philadelphia: Linguistic Data Consortium, 1999.

[29] Y. LeCun and Y. Bengio, *Convolutional networks for images, speech, and time series*, Arbib, Michael A. (ed.). The handbook of brain theory and neural networks (Second ed.), The MIT press, 255-258, 1998.

[30] S. Madikeri, I. Himawan, P. Motlicek and M. Ferras, *Integrating Online I-vector extractor with Information Bottleneck based Speaker Diarization system*, INTERSPEECH, Dresden, Germany, 2015.

[31] V. Peddinti, D. Povey, and S. Khudanpur, *A time delay neural network architecture for efficient modeling of long temporal contexts*, in INTERSPEECH, Dresden, Germany, 2015.

[32] S. Wang, Z. Huang, Y. Qian and K. Yu, *Deep Discriminant Analysis for i-vector Based Robust Speaker Recognition*, International symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, 2018.

[33] Y. Wen, K. Zhang, Z. Li and Y. Qiao, *A Discriminative Feature Learning Approach for Deep Face Recognition*, European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science, 9911, Springer, Cham, 2016.

[34] N. Ryant, M. Liberman and J. Yuan, *Speech Activity Detection on YouTube Using Deep Neural Networks*, INTERSPEECH, Lyon, France, 2013.

[35] D. Ververidis, C. Kotropoulos, *A State of the Art Review on Emotional Speech Databases*, Australian Orthodontic Congress, 2003.

[36] R. Fernandez, R.W. Picard, *Modeling Driver's Speech Under Stress*, ISCA Workshop on Speech and Emotions, Belfast, 2000.

[37] B. Schuller, B.V. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, *Acoustic emotion recognition: A benchmark comparison of performances*, IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 25(2), 233–240, 2009.

[38] K.R. Scherer, *Emotion effects on voice and speech: Paradigms and approaches to evaluation*, ISCA Workshop on Speech and Emotions, Belfast, 2000.

[39] J.H.L. Hansen and S.E. Bou-Ghazale, *Getting started with SUSAS: a speech under simulated and actual stress database*, EUROSPEECH, 1743-1746, 1997.

[40] M. El Ayadi, M.S. Kamel, and F. Karray, *Survey on speech emotion recognition: Features, classification schemes, and databases*, Pattern Recognition, 44(3), 572-587, 2011.

[41] M. El Ayadi, M.S. Kamel, F. Karray, *Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence*, Applied Science, 9(11), 2298, 2018.

[42] G. Senthil Raja and S. Dandapat, *Speaker recognition under stressed condition*, Int J Speech Technol, 13, 141–161, 2010.

[43] A. Mansour, F. Chenchah, and Z. Lachiri, *Emotional speaker recognition in real life conditions using multiple descriptors and i-vector speaker modeling technique*, Multimed Tools Appl, 78, 6441–6458, 2019.

[44] B. Schuller, J. Stadermann, and G. Rigoll, *Affect-Robust Speech Recognition by Dynamic Emotional Adaptation*, International Conference on Speech Prosody, Dresden, Germany, 2006.

[45] S. Shukla, S. Dandapat, and S.R. Mahadeva Prasanna, *A Subspace Projection Approach for Analysis of Speech Under Stressed Condition*, Circuits, Systems, and Signal Processing, 35, 4489-4500, 2016.

[46] B.H. Prasetio, H. Tamura, and K. Tanno, *The long short-term memory based on i-vector extraction for conversational speech gender identification approach*, Artif Life Robotics, 25(2), 233–240, 2020.