Ensemble Based Speaker Verification Using Adapted Score Fusion in Noisy Reverberant Environments

Ryosuke Nakanishi*, Sayaka Shiota* and Hitoshi Kiya*
* Department of Information and Communication Systems,
Tokyo Metropolitan University, Hino, Tokyo 191-0065, Japan

***Abstract*—This paper proposes an ensemble based automatic speaker recognition (ASV) using adapted score fusion in noisy reverberant environment. It is well known that background noise and reverberation affect the performance of the ASV systems. Various techniques have been reported to improve the robustness against noise and reverberation, and an ensemble based method is one of the effective techniques in the noisy environment. The ensemble based method uses a combination of several weak learners to achieve higher performance than a single learner method. However, since the performance is depended on the fusion weights, the adequate weight estimation method is required. The proposed weight estimation method is based a supervised adaptation and the evolutionary update algorithm. The QUT-NOISE-SRE protocol, which has been published recently, is used for simulating the reverberation of the clean speech in our experiments. The experimental results report the characteristics of the QUT-NOISE-SRE protocol and the effectiveness of the proposed method in noisy reverberant environment.**

I. Introduction

The performances of the state-of-the-art automatic speaker verification (ASV) systems in clean condition have achieved impressive levels in recent years (e.g., i-vector [1], PLDA [2]–[4]). Nevertheless, when the input signals are affected by noisy and reverberant environment, the performances degrade drastically. Over the years, several noise robust algorithms and techniques for ASV systems have been proposed [5]–[9]. When the noise condition of the training data is same as that of the test data (the matched condition case), the ASV system is able to keep the performance. However, it is difficult to know the noise condition of the test data. One of the effective techniques is a model selection approach. The first process of the model selection approach is noise estimation of the test data. Then, the ASV system is able to select the most appropriate model for the test data. The model selection approach is effective for the simple noise pattern. However, the real noise condition is more complicated and the ASV system is required to represent more high complexity models.

The ensemble learning is a type of machine learning that applies a combination of several weak learners to achieve an improved performance than a single learner. Bagging [10] and AdaBoost [11] are standard ensemble learning techniques, and these approaches are adopted to the ASV systems. The ensemble based methods are able to represent the complicate model for the ASV systems [12], [13]. However, the performance of the ensemble speaker verification (ESV) systems are depended on the fusion weights, the adequate weight estimation method is required.

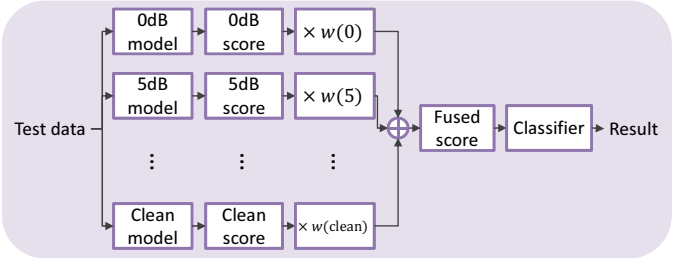


Fig. 1: The flow of score fusion using speaker dependent (SD) models estimated at each SNR.

To estimate the adequate fusion weights, this paper proposes an supervised adaptation and evolutionary adaptation algorithm. In this algorithm, the fusion weights are updated only when the equal error rate (EER) of the development set become better. The QUT-NOISE-SRE protocol [14] has been published recently for simulating the reverberation. In the experiment part, at first, the performance of the ASV system against noisy and reverberation environment by using QUT-NOISE-SRE protocol is explored. Second, the performance of the proposed method is evaluated in noisy reverberant environment.

The rest of this paper is organized as follows. Section 2 briefly describes an ensemble speaker verification system. In Section 3, the weight adaptation and upload algorithm are shown. The experimental conditions and results are illustrated in Section 4. Concluding remarks and future work are presented in Section 5.

II. Ensemble speaker verification

Conventional automatic speaker verification (ASV) systems use a single classifier to calculate the speaker similarity of an input speech. While the performances of the systems are drastically improved by the several conventional systems [1]–[4], the ASV systems are required to obtain more high performances especially for noisy environments. It have been reported that the ensemble speaker verification (ESV) systems contribute to represent the complex structure of the speaker and noise environments [15], [16]. And, the over-fitting problem can be relaxed by the fusion of ensemble classification

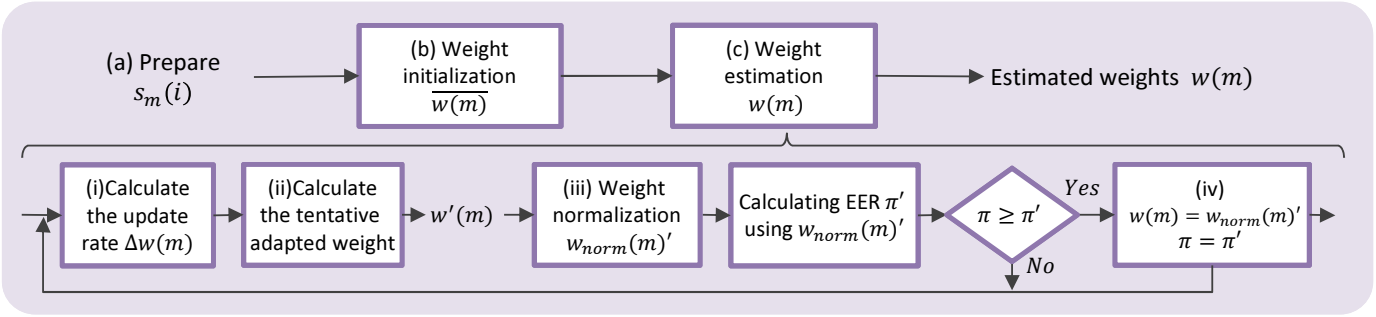


Fig. 2: The algorithm of the proposed estimation method.

systems as well [17], [18]. Bagging [10] and AdaBoost [11] are standard ensemble learning techniques, which the several models can be combined as weak classifiers and a decision can be taken relatively fast since the classifier is a simple weighted linear combination of outputs. These techniques are adopted to realize the ESV systems. However, since the performance of ESV system is dependent on the fusion weights. Thus, this paper proposes the estimation algorithms of the fusion weights for noisy reverberant environments by the evolutionary adaptation algorithm [19].

III. Proposed Method

Figure 1 shows the classification flow of the proposed method. In this ensemble framework, each weak classification is represented as the set of the speaker dependent models with each noise level. The final score is combined from each SNR model by using the adapted fusion weights. To estimate the appropriate fusion weights, the evolutionary adaptation algorithm is used. Figure 2 illustrates the proposed algorithm. The procedure is divided into three parts: (a) preparation, (b) weight initialization and (c) weight estimation. In the preparation part (a), the initial classification score $s_m(i)$ is calculated with a development data i at a SNR model m. The fusion weights for each SNR model are initialized in the step (b) as the Uniform distribution:

$$\overline{w(m)} = \frac{1}{M}. \tag{1}$$

where M is the number of the weak classifications. The process of the (c) is composed of the four steps. Each step is carried out as following process: (i) calculate the update rate, (ii) calculate the tentative adapted weight, (iii) weight normalization and (iv) judge the adapted weight is effective or not. In the step (i), the update rate $\Delta w(m)$ is calculated as follows:

$$\Delta w(m) = s_m(i) \cdot \rho \cdot y(i), \tag{2}$$

where ρ stands for the weight adapted rate. $y(i) \in \{+1, -1\}$ means the label of each development data that the data is belong to enrollment speakers or imposter speakers. In the step (ii), the tentative adapted weight $w(m)'$ is calculated with the update rate $\Delta w(m)$ and present weight $w(m)$. At the first iteration, the present weight $w(m)$ is set to the initial weight $\overline{w(m)}$.

$$w(m)' = w(m) + \Delta w(m). \tag{3}$$

TABLE I: Experimental conditions

Database (UBM)	JNAS (female)
Training data (UBM)	23657 sentences × 7 SNRs (165599 sentences)
Database	VLD database
Training data (Speaker dependent model)	70 sentences × 17 speakers (1190 sentences)
Development data	10 sentences × 17 speakers (170 sentences)
Test data	20 sentences × 17 speakers (340 sentences)
# of mixtures	1024
Sampling frequency	16 kHz
Frame length / Frame shift	25 msec / 10 msec
Features	MFCC 19 order+ Δ + $\Delta\Delta$

Since the tentative adapted weight $w(m)'$ is sometimes set as a large number, the normalization process is put in the step (iii).

$$w_{norm}(m)' = \frac{w(m)'}{\sum_{k \in M} w(k)'}. \tag{4}$$

The step (iv) is the judgment process against the normalized weight $w_{norm}(m)'$. At first, the performance of the ESV system is evaluated with the development data and the normalized weight $w_{norm}(m)'$. To evaluate the performance, the Equal Error Rate (EER) is used. The EER π' calculated from the normalized weight $w_{norm}(m)'$ weight $w(m)'$ is compared with the EER π calculated from the present weight $w(m)$. When $\pi' \leq \pi$, the weight is uploaded:

$$w(m) = w_{norm}(m)'. \tag{5}$$

The weight update steps are repeated, and the final estimated weight $w(m)$ is used for the ESV system as the appropriate fusion weights.

IV. Evaluation Experiment

To investigate the effectiveness of the proposed method, the speaker verification experiments are conducted. The GMM-UBM framework [20] is used as the traditional ASV system.

A. Experimental Conditions

Table I shows the experimental conditions. To evaluate the performance of the ASV systems at each SNR, a variety of the SNRs is set to $M = \{0, 5, 10, 15, 20, 30, \text{clean}\}$. The universal background model (UBM) is estimated under multi-condition training with clean speech and noisy speech at each SNR

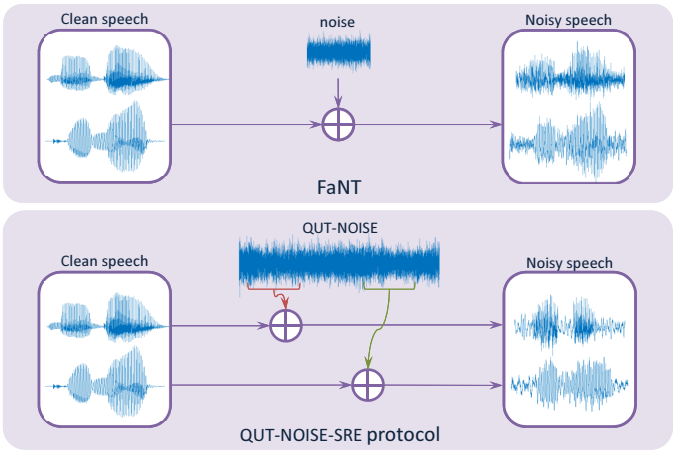


Fig. 3: The differences between the two noise convolution tools (FaNT and QUT-NOISE-SRE protocol).

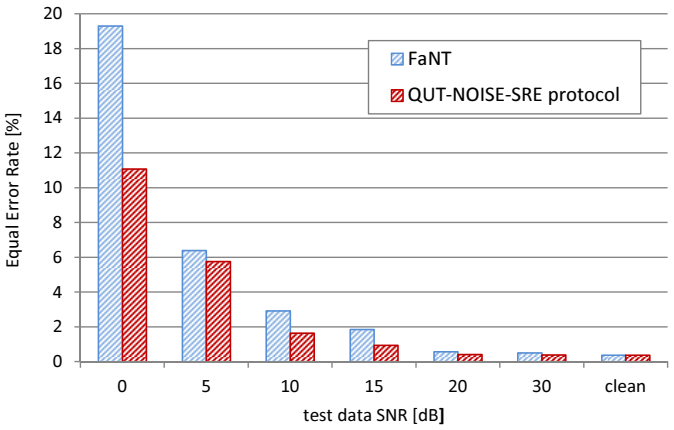


Fig. 4: The EERs at each test data SNR by the use of two noise convoluting protocol.

TABLE II: Compared methods

Methods	SD models SNR	Score fusion
MC	Model SNR is matched test data SNR	—
BC (Oracle)	Model SNR at the lowest EER against test data	—
Spectral subtraction (SS)	Clean model (verifying test data applied SS)	—
Uniform	Combining SNR models after combining score in each SNR model	The uniform weight is set in all SNR models.
Proposed method	Combining SNR models after combining score in each SNR model	Each weight is updated when the EER against development data decreases.

(0, 5, 10, 15, 20 and 30dB). All female speakers in JNAS database [21] are used for estimating the UBM. Noise of automobile cabin (car) in JEIDA noise database [22] is added to the UBM training data by Filtering and Noise Adding Tool (FaNT) [23]. Maximum a posteriori (MAP) adaptation is used to adapt the UBM for each SD model. Speech data, recorded by using the headset microphone (SHURE SM10A-CN) with a pop filter from VLD database [24], are used for the training of SD models, the development data and the test data. The number of speakers is 17 and all speakers are female. For the speech samples from the VLD database, CAR scenario in the QUT-NOISE database [25] is added by the QUT-NOISE-SRE protocol [14].

Figure 3 shows the differences between the two noise convolution tools, which are FaNT and QUT-NOISE-SRE protocol, respectively. In the FaNT procedure, the single segment is selected from the noise data to convolute the noise. On the other hand, the QUT-NOISE-SRE protocol convolutes random segments from the scenario to a clean speech.

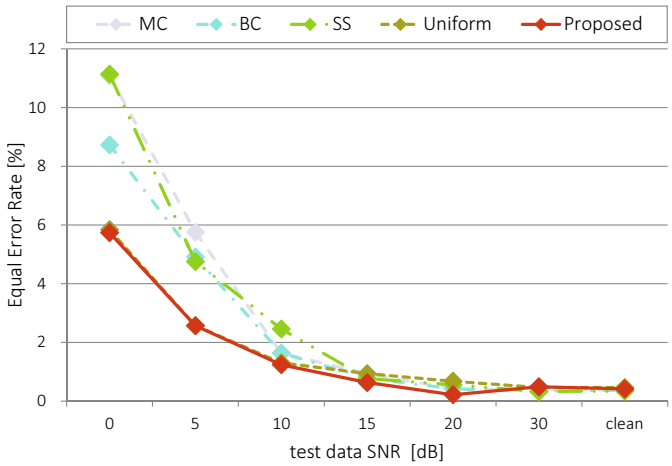
Recently, databases which are recorded in some real environments are widely used for ASV evaluation. While the effectiveness for these database is important, it is also important to use the simulated noise and reverberant for speech database and to investigate the simple effect against to the ASV systems. Thus, we evaluate the performance of the ASV systems in environments with reverberant and non-reverberant speech data, respectively. The weight adaptation rates ρ in reverberant and non-reverberant conditions are experimentally set in 0.003 and 0.005, respectively. These values are determined experimentally.

Table II illustrates the compared methods. The matched condition (MC), the best condition (BC) and the spectral subtraction (SS) [26] use a single classifier to calculate the speaker similarity. Uniform is calculate the speaker similarity by combining all SNR models with uniform weights: $w(m) = 1/7$.

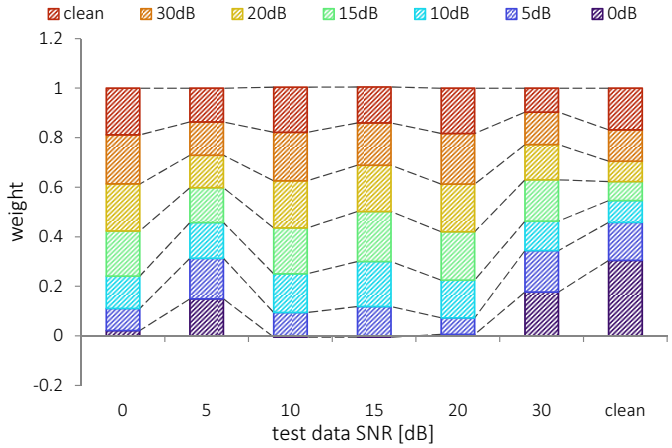
B. Experimental results

1) Comparing the noise convolution protocols: Firstly, the speaker verification experiment in MC is conducted to evaluate the difference of verification results by the FaNT and the QUT-NOISE-SRE protocol. Figure 4 shows EERs at each test data SNRs. In the all SNR conditions, the EER of the QUT-NOISE-SRE protocol is lower than the EER of the FaNT. The QUT-NOISE-SRE protocol uses several kinds of noise segment to convolute to clean speech. Thus, the training model of the QUT-NOISE-SRE protocol can be estimated as the more complicate model than that of the FaNT.

2) Non-reverberant condition: Figure 5 (a) shows the EERs at each test data SNR by each method under the noisy and non-reverberant environment. Compared ensemble methods (Uniform and the proposed method) with the single classification methods (MC, BC, SS), the ensemble methods obtained lower EERs than the single classification methods, especially the low SNR case. Even though the SS is the method of the noise reduction, the original speech is also affected and the performance is not better than ESV methods. Since the ensemble methods can be represented complicated models, the performance is improved. Comparing the proposed method with the Uniform, the EERs of the proposed method is lower than that of the Uniform. Hence, the appropriate fusion weights are affected to the performance of the ESV system in the noisy conditions. Figure 5 (b) shows the final score fusion weights that are estimated by the proposed method. Each item

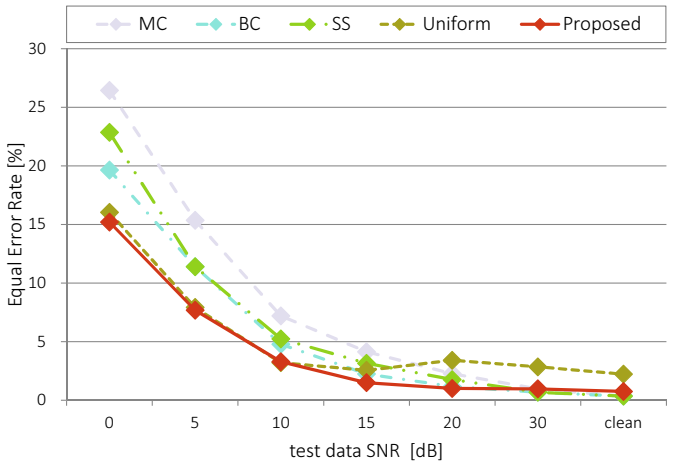


(a) The EERs at each test data SNR by the proposed method and other methods.

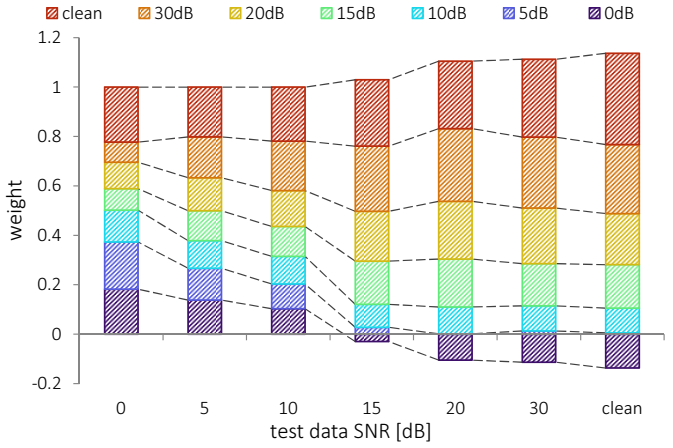


(b) The final weight by the proposed method

Fig. 5: The EERs under the non-reverberant environment and the adapted weight.



(a) The EERs at each test data SNR by the proposed method and other methods.



(b) The final weight by the proposed method

Fig. 6: The EERs under the reverberant environment and the adapted weight.

represents the final weight against the score obtained from each SNR model at each test data SNR. The horizontal and the vertical axis illustrate test data SNR and the value of the final adapted weight, respectively. From fig. 5 (b), the adapted weight varies at each test data SNR. Specifically, the weight at 0dB model changes widely.

3) Reverberant condition: Figure 6 (a) shows the EERs of each method under the noisy and reverberant environment. The SS can reduce the influence of reverberation compared to the MC and achieve the performance close to the BC. The BC and the ensemble methods have the same tendency to the case of the non-reverberant environment. The ensemble methods are effective at low SNRs of test data under the reverberant environment. Nevertheless, the performance of the Uniform is worse at high SNRs of test data under the reverberant environment. On the other hand, the proposed method provides the high performance in the noisy reverberation condition comparing with the other methods. From these results, the proposed method is also robust against under noisy reverberant environment.

Figure 6 (b) shows the score fusion weights that are finally estimated by the proposed method. It can be seen from that the EERs of the proposed method are close to those of the Uniform. The score fusion weight can be adapted so that the score fusion weights against the score obtained from high SNR models are increased at high SNRs of test data. The proposed method achieves the high performance without the degradation of the performance.

Comparing non-reverberant condition with reverberant condition, the values of the estimated weights are different. And, the EERs of the proposed method is lower than that of the Uniform. Thus, each noise condition is required the appropriate fusion weights to obtain the high performance of the ESV system. Because the proposed method is adopted the evolutionary algorithm, the performance becomes better when the enough development data can be obtained.

Moreover, an evaluation experiment using another database is also conducted under the same conditions. We obtained the almost same results as this paper.

V. Conclusion

This paper proposed an ensemble based ASV using adapted score fusion in noisy reverberant environment. The proposed weight estimation method is based a supervised adaptation and adjustable update algorithm using development data. To evaluate the robustness of the proposed method under noisy reverberant environment, the QUT-NOISE-SRE protocol is used for simulating the reverberation of the clean speech in our experiments. The experimental results showed the effectiveness of the proposed method in noisy reverberant environment.

In order to obtain more reliable results, it is necessary to increase the amount of the experimental data. Moreover, a cost function based the weight estimation method should be considered. In this paper, our evaluation experiments have been conducted by using development data selected the specific domain. However, evaluation results are considered to change if development data are selected from another domain. It should be investigated how depends on the domain of the development data in our future work.

References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV 2006*, pp. 531–542, 2006.

[3] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, 2007.

[4] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[5] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[6] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[7] M. I. Mandasari, M. McLaren, and D. A. Van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4249–4252, 2012.

[8] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 836–845, 2014.

[9] N. Thatphithakkul, B. Kruatrachue, C. Wutiwiwatchai, S. Marukatat, and V. Boonpiam, "Tree-structured model selection and simulated-data adaptation for environmental and speaker robust speech recognition," in *Communications and Information Technologies, 2007. ISCIT'07. International Symposium on*. IEEE, 2007, pp. 1570–1574.

[10] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[11] J. Friedman, T. Hastie, R. Tibshirani *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[12] A. Park and T. J. Hazen, "ASR dependent techniques for speaker identification." *INTERSPEECH*, 2002.

[13] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, 2012.

[14] D. B. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," *Proceedings of the 16th Annual Conference of the International Speech Communication Association, Interspeech 2015*, pp. 3456–3460, 2015.

[15] Y. Tsao and C.-H. Lee, "An ensemble speaker and speaking environment modeling approach to robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 1025–1037, 2009.

[16] Y. Tsao, S. Matsuda, C. Hori, H. Kashioka, and C.-H. Lee, "A MAP-based online estimation approach to ensemble speaker and speaking environment modeling," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 2, pp. 403–416, 2014.

[17] V. Hautamäki, K.-A. Lee, A. Larcher, T. Kinnunen, B. Ma, and H. Li, "Variational bayes logistic regression as regularized fusion for NIST SRE 2010." *Odyssey*, pp. 268–274, 2012.

[18] V. Hautamaki, T. Kinnunen, F. Sedlák, K. A. Lee, B. Ma, and H. Li, "Sparse classifier fusion for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 1622–1631, 2013.

[19] T. Shinozaki and S. Watanabe, "Structure discovery of deep neural network based on evolutionary algorithms," *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4979–4983, April 2015.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[21] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research." *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[22] S. Itahashi, "A noise database and Japanese common speech data corpus," *J. Acoust. Soc. Jpn*, vol. 47, no. 12, pp. 951–953, 1991.

[23] H.-G. Hirsch, "F a NT-Filtering and Noise Adding Tool," 2005.

[24] S. Shiota, V. Fernando, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification," *Proc. Interspeech*, pp. 239–243, 2015.

[25] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," *Proceedings of Interspeech 2010*, 2010.

[26] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.