# Towards Noise-Robust Automatic Speaker Verification Using Pop Noise

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

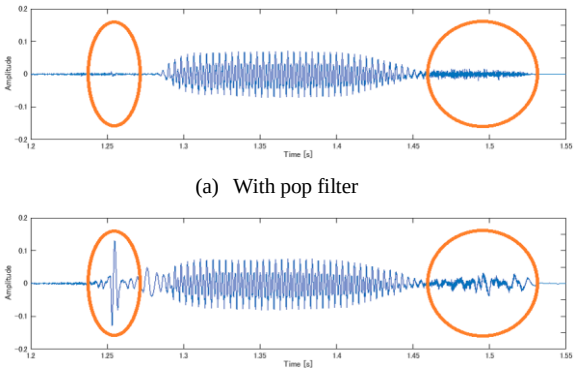
***Abstract*—This paper proposes a novel framework to actively use pop noise, which is unconsciously caused by human breath, for noise-robust automatic speaker verification (ASV). Recently, supervector-based or i-vector-based approaches have been achieved significant improvements for robust ASV systems. However, the performance of the approaches is generally degraded in noisy environments, when types of noise and/or signal-to-noise ratio (SNR) are unknown in particular. This paper focuses on the use of pop noise that is expected to include the information on both noise signals and speaker characteristics. To evaluate the effectiveness of the proposed framework, ASV experiments under typical noise environments are conducted. The results show that the proposed ASV framework with pop noise can improve the performance even though the noise conditions of the training data, e.g. types of noise and/or SNRs, are different from those of the test data.**

***Keywords—automatic speaker verification; noise robustness; pop noise; UBM-GMM***

## I. INTRODUCTION

Automatic speaker verification (ASV) which uses only speaker's voice samples is well known as a useful biometric authentication approach. Recently, the performance of ASV systems has been improved by the i-vector [1] or PLDA [2-4] approach, and a lot of articles regarding the state-of-the-art schemes have been published to show the potential to support mass-market adoption [5]. However, the performance of the ASV systems in real commercial and forensic applications is degraded by the influence of background noise. Thus, one of the most important challenges in the area of ASV is to make the system robust towards its acoustic environment [6-12]. Many approaches dealing with noise robustness have been researched [13, 14], such as spectral subtraction [15, 16] and Wiener filter [17-19]. However, when types of noise and/or signal-to-noise ratio (SNR) are unknown in particular, the performances of the ASV systems are still insufficient [20, 21].

Since pop noise in speech is a well-known distortion, occurring when human breath reaches directly a microphone, pop noise suppression techniques have been proposed to reduce its influence. On the other hand, statistical model based speech synthesis or voice conversion systems have considered that breathing contains some specific information of a speaker [22-24]. Moreover, it has been reported that the performance of ASV system is improved by using pop noise actively [25]. This paper has been inspired by this insight that pop noise has the information of both a speaker and non-stationary noise. Thus, this paper focuses on the use of pop noise to achieve noise-robust ASV systems.




(a) With pop filter

(b) Without pop filter

Fig. 1. Waveforms recorded by using two headset microphones. Only the waveform recorded without any pop filter has pop noise.

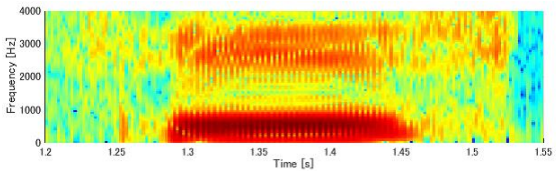
To evaluate the robustness of the proposed framework, a new database including pop noise is constructed. A model with pop noise is estimated by using the database. Using the new database, it is observed that the equal error rate (EER) of the proposed model is 15.1%, which is lower than that of the noise matched condition models at 0dB SNR. Comparing the proposed model with conventional noise reduction methods, the proposed model provides good EERs at several SNRs. From these results, the pop noise is able to contribute the noise-robustness of the ASV systems.

This paper organizes as follows: In section 2, the characteristics of pop noise are described, and the details of the designed database are illustrated in section 3. Section 4 shows experimental results, and finally section 5 gives our conclusions and future work.

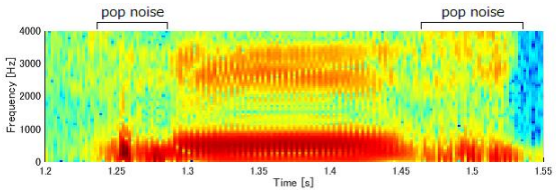
## II. CHARACTERISTICS OF POP NOISE

### A. *Speaker individuality of pop noise*

Characteristics of spontaneous speech can be classified into




(a) With pop filter




(b) With pop filter

Fig. 2. Spectrograms of a speech recorded by using two headset microphones.

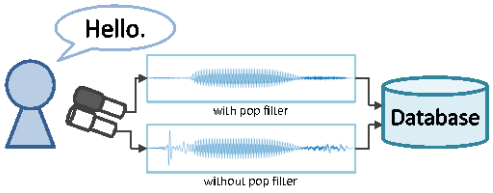



Fig. 3. The flow of recording a speech.

Fig. 4. Headset microphone with a pop filter (right) and that without any pop filter (left).

paralinguistic cues, disfluency patterns and reflection phenomena [26]. real speech, but are always present in spontaneous speech. Especially, breathing could contain speaker's some specific information. *Fig. 1* shows the difference of waveforms that were simultaneously recorded by two headset microphones, where one was with a pop filter and the other was without any pop filter. The pop filter usually serves to reduce the influence of breathing [26, 27], and hence the waveform in *Fig. 1* (b) includes a distortion due to the influence of breathing. The distortion has been conventionally regarded as a noise and suppressed from speech signals. However, statistical based speech synthesis and voice conversion systems have considered that breathing contains specific information of a speaker [23, 24, 26]. Moreover, it has been reported that the performance of ASV systems is improved by using pop noise actively [25].

### B. *Noise robustness using pop noise*

A noisy signal $\boldsymbol{Y}$ is represented as the resultant vector of a speech vector $\boldsymbol{X}$ and a background noise vector $\boldsymbol{N}$ :

$$\boldsymbol{Y}^{(s)} = \boldsymbol{X}^{(s)} + \boldsymbol{N}, \quad (1)$$

where $s$ denotes a speaker ID. By contrast, a speech signal $\boldsymbol{Y}$ including pop noise is represented as the resultant vector of a speech vector $\boldsymbol{X}$ and a pop noise as below:

$$\boldsymbol{Y}^{(s)} = \boldsymbol{X}^{(s)} + \boldsymbol{PN}^{(s)}, \quad (2)$$

Since the phenomenon of the pop noise vector $\boldsymbol{PN}$ depends on each speaker, $\boldsymbol{PN}$ has the information of both a speaker and non-stationary noise. Thus, this paper focuses on the use of pop noise to achieve noise-robust ASV systems. *Fig. 2* compares the spectrogram of a recording with a pop filter to that of a recording without any pop filter. It can be seen that the energy of very low-frequency is strongly affected by pop noise. Thus, the speaker dependent (SD) model estimated with pop noise is able to capture low-frequency components such as environmental noises, and thereby the noise robustness of the ASV system with pop noise is improved.

## III. Speech Database with Pop Noise

Since the standard speech database such as the NIST database [28] has not been recorded any speeches including pop noise,we design a new speech database including pop noise. In order to design the database with pop noise, two kinds of microphones, in which one is with a pop filter and the other is without any pop filter, are simultaneously used to record speeches. *Fig. 3* shows the flow of recording a speech via the two microphones. We use the headset microphones (SHURE SM10A-CN) for designing the database shown in *Fig. 4* Since the headset microphones are supposed to be used in the close distance to speaker's mouth, they are generally designed to have robustness of breathing. Nevertheless, the speech of the headset microphones without a pop filter includes some pop noise as shown in *Fig. 2* and *3*. The number of speakers is 17 and all speakers are female. Each speaker speaks 100~sentences. Half of the sentences are common sentences and the other half are selected at random from Japanese Newspaper Article Sentences (JNAS) [29], which is one of the standard database for speaker and speech recognition research area in Japan. The common part is used to measure the balance of phonemes, and the other part is restricted to only short sentences.

Noises used in our experiments are from the JEIDA noise database [30]. Two kinds of noise samples, i.e. automobile cabin (car) and exhibition hall, are used from JEIDA. The car noise is more stable than the exhibition hall one, and the exhibition hall noise includes some voices. The noisy speech samples are prepared by adding selected noise samples to recorded samples at various SNRs (0, 5, 10, 15, 20 and 30dB). The artificially distorted speech samples are created by using Filtering and Adding Noise Tool (FaNT) [31].

## IV. Evaluation Experiment

To evaluate the noise robustness of pop noise, the speaker verification experiments are conducted. All SD models are estimated by using UBM-GMM framework [32].

TABLE I. EXPERIMENTAL CONDITIONS

| | |
|---|---|
| Training data (Speaker dependent model) | 70 sentences × 17 speakers (1190 sentences) |
| Test data | 30 sentences × 17 speakers (510 sentences) |
| Database for UBM | JNAS (female) |
| Training data for UBM | 23657 sentences |
| GMM mixtures | 1024 |
| Sampling freq. | 16 kHz |
| Bit rate | 16 |
| Frame length | 25 ms |
| Frame shift | 10 ms |
| Features | MFCC 19 order + $\Delta$ + $\Delta\Delta$ |

### A. *Experimental conditions*

Table 1 shows the experimental conditions. The database described in section 3 is used for estimating SD models. The database has two channels for every sentence, and thus each channel is represented by the combination of an alphabet and a number; such as "H0" or "H1". The alphabet "H" means the headset microphone, and the number "1" or "0" indicates whether a pop filter is used or not. The "proposed model" stands for SD models estimated with pop noise (H0 data).

When the noise level and type are previously known, the SD model is able to be estimated with the matched condition to test data, and the performance is easily improved (MC model). In the first experiment, the matched condition models estimated with H0 channel data (MC(H0)) and with H1 channel data (MC(H1)) are compared with the proposed model.

To realize reliable ASV systems under noisy environments, the use of noise reduction technique becomes one of the solutions. Therefore, the second experiment investigates whether the proposed models provide better effect on the noise robustness than the noise suppression techniques or not. For this experiment, the Wiener filter based noise reduction methods, the decision-directed (DD) [18] and the Two-Step Noise Reduction (TSNR) [19] methods, are applied to the test data. It has been reported that these algorithms can estimate reliable SNR by preserving the onset and offset of a speech [33, 34].

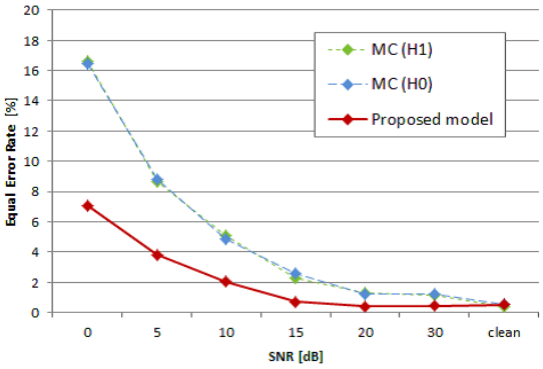
To calculate equal error rates (EERs) for every experiment, Z-Norm [35] is used as one of standard score normalization methods for likelihood ratio scoring below:

$$S_{ZN}(\boldsymbol{X}, \boldsymbol{\lambda}) = \frac{S(\boldsymbol{X}, \boldsymbol{\lambda}) - \mu_I}{\sigma_I}, \qquad (3)$$

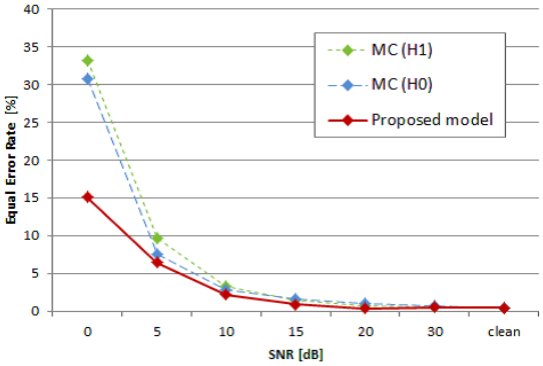
where $S(\boldsymbol{X}, \boldsymbol{\lambda})$ denotes the likelihood ratio score derived from the input data $\boldsymbol{X}$, the SD model $\boldsymbol{\lambda}_U$ and the imposter speaker model $\boldsymbol{\lambda}_I$. $\mu_I$ and $\sigma_I$ represent the mean and variance of the likelihood ratio score against imposter speech.

### B. *Experiment results*

*Fig. 5* (a) and (b) show EERs at each SNR under the car noise and the exhibition hall noise, respectively. Clean represents that no noise is added to the test data. Compared the proposed model with MC(H1), the EERs of the proposed model are lower than those of MC(H1) at all SNRs under the car noise. The EERs of the proposed model are also lower than




(a) Automobile cabin (car) noise

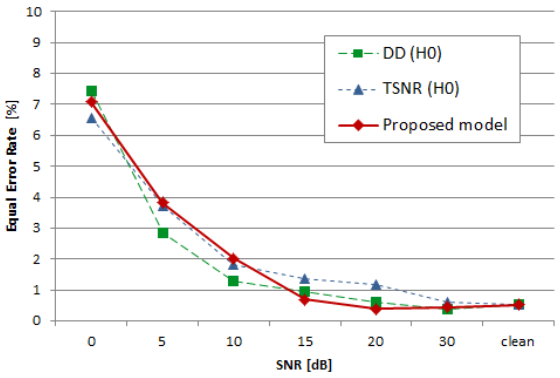



(b) Exhibition hall noise

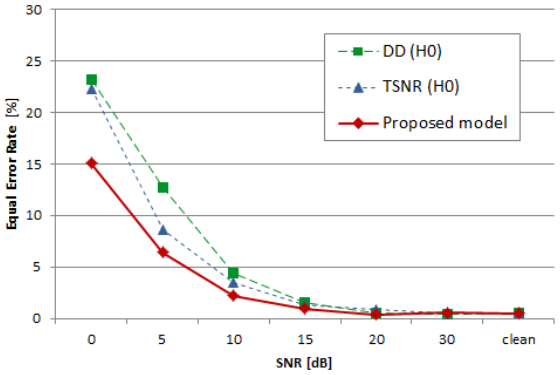
Fig. 5. Experimental results of comparing the matched condition models with the proposed model.

those of MC(H0). These results indicate that the proposed model considering pop noise can improve the noise robustness, without estimating environmental noise. The EERs of the proposed model in the exhibition hall have almost the same as those in the car noise. These results illustrate that the SD model estimated with pop noise has robustness against background noise, even though the noise is unknown. On the other hand, the EERs of MC(H1) are almost same as those of MC(H0) in both the car noise and the exhibition hall one. Even though MC(H0) model includes some pop noise features, the EERs of MC(H0) obtained no improvement. It can be considered that the pop noise for MC(H0) is treated as with background noise. From these results, when the pop noise signal is specifically different from the speech signal and the background noise, the pop noise components in the SD model are able to behave like accepting noise signals. Since the phenomenon of pop noise is regarded as a non-stationary noise, the components of pop noise are represented as a complex noise distribution. Hence, the proposed model can improve the performance under several SNR levels and types of background noise.

*Fig. 6* (a) and (b) denote the EERs of the proposed model and the noise reduction methods (DD and TSNR). From *Fig. 6* (a), the three methods have almost same EERs under the car noise environment. The Wiener filter based noise reduction methods, i.e. the DD and TSNR methods, are applied to the test data, where the test data for the proposed model is without




(a) Automobile cabin (car) noise



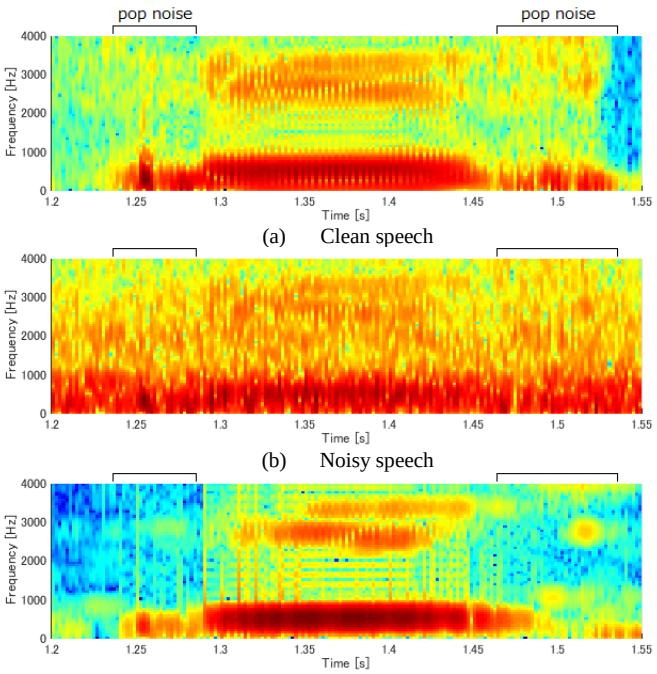

(b) Exhibition hall noise

Fig. 6. Experimental results of comparing the noise reduction method by Wiener filtering with the proposed model.

any kind of noise reduction methods. This result indicates that the proposed model can reduce some noise influences as a noise reduction method. Additionally, in the exhibition hall noise (*Fig. 7* (b)), the EER of the proposed model achieves 15.1% at 0dB SNR. Since the error reduction rates from the DD and TSNR methods are respectively 35.0% and 32.5%, the use of pop noise at low SNRs improves the ASV performance very well. The original speech data is deteriorated with a noise reduction filter, and the filter may degrade the ASV performance. On the other hand, the proposed model can actually evaluate the original test data without any filter, and thus the ASV system using pop noise can achieve this improvement. *Fig. 7* shows the spectrograms of a clean speech, the noisy speech and the speech filtered by the DD method. It is observed that the pop noise and speech spectrograms of *Fig. 7* are degraded by the DD method based filtering. It is considered that the performance of the ASV system is able to effectively improve by using pop noise.

## V. Conclusion

This paper focused on the use of pop noise to achieve noise-robust ASV systems. To investigate the effectiveness, the ASV system using pop noise was constructed. A new database including pop noise in speech signals was designed and the evaluation tests under the car and exhibition hall noises




(a) Clean speech

(b) Noisy speech

(c) Filter processed noisy speech

Fig. 7. Spectrograms of a clean speech, the noisy speech and the speech filtered by the DD method without any pop filter. The noisy speech is prepared by adding the exhibition hall noise at 0dB SNR. Filtering a speech with the DD method reduces the effects of pop noise.

were carried out. The experimental results showed that the use of pop noise improves the noise robustness without estimating models for specific noise condition or using noise reduction methods. Only UBM-GMM experiments were carried out for the preliminary investigation in this paper. However, the state-of-the-art ASV systems, such as i-vector or PLDA, should also be carried out to show the effectiveness of pop noise in the near future. We expect that the effectiveness of pop noise may become less but obtain better results because the i-vector framework is based on UBM-GMM models. Since the new database was designed for preliminary investigations, the amount of data will be also increased for 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," in Computer Vision–ECCV 2006. Springer, 2006, pp. 531–542.

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

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

[5] B. G. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 7, pp. 1960–1968, 2007.

[6] 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.

[7] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 1. IEEE, 1998, pp. 121–124.

[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] T. Hasan and J. H. Hansen, "Acoustic factor analysis based universal background model for robust speaker verification in noise."in INTERSPEECH, 2013, pp. 3127–3131.

[10] S. Sarkar and K. S. Rao, "A novel boosting algorithm for improved i-vector based speaker verification in noisy environments,"in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[11] L. F. Gallardo, M. Wagner, and S. M¨oller, "I-vector speaker verification based on phonetic information under transmission channel effects," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[12] M. Man-Wai, "SNR-dependent mixture of PLDA for noise robust speaker verification," 2014.

[13] 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.

[14] T. F. Zheng, Q. Jin, L. Li, J. Wang, and F. Bie, "An overview of robustness related issues in speaker recognition," APSIPA ASC 2014, Dec 2014.

[15] 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.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79., vol. 4. IEEE, 1979, pp. 208–211.

[17] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604, 1979.

[18] P. Scalart et al., "Speech enhancement based on a priori signal to noise estimation," in Proc. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 2. IEEE, 1996, pp. 629–632.

[19] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 6, pp. 2098–2108, 2006.

[20] A. E. Rosenberg, "Recent research in automatic speaker recognition,"Advances in Speech Signal Processing, vol. 5, pp. 701–738, 1992.

[21] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, vol. 2. IEEE, 2003, pp. II–53.

[22] H. Meirong, J. Huimin, and Y. Yangrui, "Study on the rhythm of tibetan poems based on breathing signal," in Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on, vol. 3. IEEE, 2010, pp. 618–621.

[23] K. I. Nordstrom, G. A. Rutledge, and P. F. Driessen, "Using voice conversion as a paradigm for analyzing breathy singing voices," in Proc. Communications, Computers and signal Processing, 2005. PACRIM. 2005 IEEE Pacific Rim Conference on. IEEE, 2005, pp. 428–431.

[24] T. Nakano, M. Goto, J. Ogata, and H. Yuzuru, "Acoustic characteristics of breath sounds in solo vocal and their application to automatic breath detection." in IPSJ SIG Notes, vol. 2008, no. 78, 2008, pp. 83–88.

[25] S. Shiota, T. Matsui, and H. Kiya, "Non-acoustical noise for speaker verification," Autumn Meeting of Acoustical Society of Japan, pp. 88–89, 2014.

[26] S. Sundaram and S. Narayanan, "An empirical text transformation method for spontaneous speech synthesizers." in Proc. Interspeech. Citeseer, 2003.

[27] L. C. Oliveira, S. Paulo, L. Figueira, C. Mendes, A. Nunes, and J. Godinho, "Methodologies for designing and recording speech databases for corpus based synthesis." in Proc. the Sixth Intrnational Language Resources and Evaluation (LREC), 2008.

[28] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation–overview, methodology, systems, results, perspective," Speech Communication, vol. 31, no. 2, pp. 225–254, 2000.

[29] 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.

[30] S. Itahashi, "A noise database and japanese common speech data corpus," J. Acoust. Soc. Jpn, vol. 47, no. 12, pp. 951.953, 1991.

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

[32] 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.

[33] H. Ding, Y. Soon, S. N. Koh, and C. K. Yeo, "A spectral filtering method based on hybrid wiener filters for speech enhancement,"Speech Communication, vol. 51, no. 3, pp. 259.267, 2009.

[34] S. Ou, C. Geng, and Y. Gao, "Improved a priori SNR estimation for speech enhancement incorporating speech distortion component,"TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 11, no. 9, pp. 5359.5364, 2013.

[35] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in Proc. Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, vol. 2. IEEE, 2003, pp. II.49.