VOICE LIVENESS DETECTION FOR SPEAKER VERIFICATION BASED ON A TANDEM SINGLE/DOUBLE-CHANNEL POP NOISE DETECTOR

Sayaka Shiota[1], *Fernando Villavicencio*[2], *Junichi Yamagishi*[2], *Nobutaka Ono*[2], *Isao Echizen*[2], *Tomoko Matsui*[3]

[1]Tokyo Metropolitan University, Hino, Tokyo, 191-0065, Japan.
[2]National Institute of Informatics, Chiyoda, Tokyo, 101-8430, Japan.
[3]The Institue of Statistical and Mathematics, Tachikawa, Tokyo, 190-8562, Japan.

Abstract

This paper presents an algorithm for detecting spoofing attacks against automatic speaker verification (ASV) systems. While such systems now have performances comparable to those of other biometric modalities, spoofing techniques used against them have progressed drastically. Several techniques can be used to generate spoofing materials (e.g., speech synthesis and voice conversion techniques), and detecting them only on the basis of differences at an acoustic speaker modeling level is a challenging task. Moreover, differences between "live" and artificially generated material are expected to gradually decrease in the near future due to advances in synthesis technologies. A previously proposed "voice liveness" detection framework aimed at validating whether speech signals were generated by a person or artificially created uses elementary algorithms to detect pop noise. Detection is taken as evidence of liveness. A more advanced detection algorithm has now been developed that combines single- and double-channel pop noise detection. Experiments demonstrated that this tandem algorithm detects pop noise more effectively: the detection error rate was up to 80% less that those achieved with the elementary algorithms.

1. Introduction

Biometric authentication plays an important role in reliable management systems [1, 2]. Automatic speaker verification (ASV) is an easy-to-use biometric authentication system that uses only speaker's voice samples. Its performance has been improved by making use of techniques based on i-vectors [3] or probabilistic linear discriminant analysis (PLDA) [4]. Moreover, the current performance of state-of-the-art ASV schemes makes them ready for mass-market adoption.

At the same time, there has been significant progress in speech synthesis technologies such as text-to-speech (TTS) systems [5, 6] and voice transformation or conversion systems [7]. Such system can now generate natural-sounding artificial speech for a target speaker from text or the waveform of speech uttered by someone else. Although there has been much research on these technologies for use in various applications (e.g., for assisting individuals with vocal disabilities), they can also be used for vocal identity falsification such as in spoofing attacks against ASV systems, representing a serious personal security issue [8, 9, 10]. This has led to the recent emergence of research on the definition and development of countermeasures for detecting spoofing attacks [11, 12, 13, 14]. Typically, three different types of these attacks are considered: replay, speech synthesis, and voice conversion. The countermeasure strategies are mainly based on comparing the acoustic features of artificial signals with those of natural ones [15, 16, 17], with spectral, F0, and modulation-related information used as the basis of their computation [18]. However, the acoustic differences between artificial and natural speech are expected to gradually become smaller and eventually negligible in the near future.

Looking at other biometrics fields, we see that face, fingerprint, and even iris recognition systems also suffer from spoofing attacks, and researchers are continuing to develop appropriate countermeasures [19, 20, 21]. One of the most effective countermeasures is to use a "liveness detection" framework to determine whether the attempted authentication is from an actual person (live voice). The liveness detection framework has been reported to reduced vulnerability significantly in various image processing fields [22, 23, 24].

To determine whether the presented signals originated from an actual person, their liveness needs to be evaluated. One way to do this is to detect pop noise, and several algorithms for detecting it have been reported [25]. Since pop noise is a common distortion in speech that occurs when a speaker's breath reaches the microphone and is poorly reproduced by loudspeakers [26, 27], it seems reasonable to consider it as evidence of liveness at the input of an authentication system. A measure that takes into account the presence of pop noise phenomena might therefore be well suited as the basis for discriminating between *live* and *played* speech (though loudspeakers).

We previously proposed a strategy for ASV based on the "liveness detection" framework and defined techniques for countermeasures based on voice liveness detection (VLD) with the aim of detecting spoofing materials more robustly [25]. These countermeasures include pop noise detection algorithms. More precisely, two algorithms based on two different strategies were presented. Testing showed that each has promising detection performance. To achieve even better performance, we have now integrated them into a tandem algorithm. Experimental evaluation on ASV tasks showed that tandem approach does improve performance (equal error rate from 4.73% to 0.95%).

In section 2 we briefly describe voice liveness detection. Our proposed tandem pop noise detector is presented in section 3. Section 4 presents the evaluation results, and section 5 summarizes the key points and mentions future work.

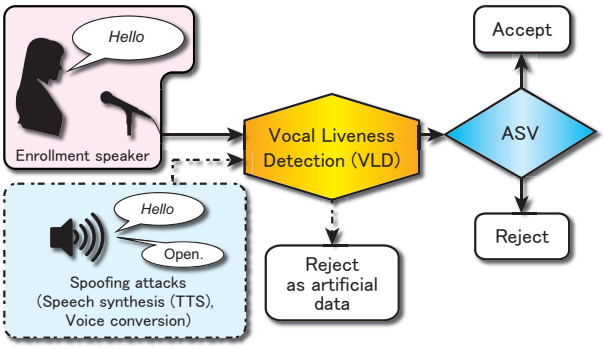

Figure 1: Overview of automatic speaker verification system including VLD module

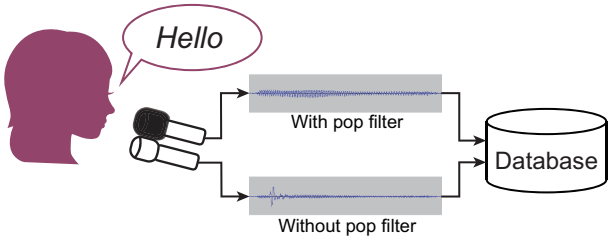

Figure 2: Recording process in double-channel algorithm

2. Voice liveness detection

2.1. Attacks on speaker verification systems

The potential for ASV to be spoofed is well recognized and there is growing interest in assessing the vulnerabilities of ASV systems and developing robust countermeasures against spoofing attacks [8, 9]. There are three main types of spoofing attacks: replay, speech synthesis, and voice conversion. Each type of attack is defined as follows:

- Replay: replay of pre-recorded utterances of the target person.
- Speech synthesis: automatic generation of natural-sounding artificial speech for a target person from text.
- Voice conversion: conversion of attacker's natural voice into that of targeted person.

Several countermeasures against each type of spoofing attack have been reported. For replay attacks, we can simply use text prompting and change the prompts every time [28, 29]. However, for spoofing attacks based on material generated by means of speech synthesis and voice conversion techniques, none of the reported countermeasures provide a fundamental solution [30]. Considering the actual potential scenarios for spoofing attacks, we can assume that they are based on replaying the spoofing material, through loudspeakers. Accordingly, independently of the nature of the spoofing material, our task is to basically discriminate between speech produced by an actual person and speech played through loudspeakers.

2.2. Framework of voice liveness detection

Figure 1 shows a diagram of an automatic speaker verification system including the VLD module. The VLD module is designed to reject all speech signals that do not show evidence of liveness regardless of the nature of the spoofing attack. Speaker verification is conducted as usual in a subsequent module. Although this figure illustrates a sequential combination of VLD and ASV modules, the VLD and ASV tasks can be integrated to work simultaneously.

As liveness evidence the, VLD detects speech waveforms characteristics unique to speech produced by an actual person. Briefly, the human voice is the result of acoustic shaping in the vocal tract of the airflow produced by interactions between the lungs and vocal chords. The resulting airflow is transformed into an acoustical signal when it is captured by a microphone. Spontaneous strong breathing can cause convolution between the airflow and vocal cavities, producing a sort of perceived plosive burst, commonly known as pop noise, which can be captured by a microphone. The acoustic conditions change when this noise is played through loudspeakers, commonly resulting in poor reproduction of the pop noise. Thus, by detecting pop noise events, we may be able to distinguish between an actual human voice and played back through loudspeakers.

3. TANDEM SINGLE-DOUBLE CHANNEL POP NOISE DETECTOR

Here we first describe the concept and process of our two previously reported algorithms [25] for capturing pop noise as liveness evidence and then introduce our tandem algorithm.

3.1. Single- and double-channel pop noise detection algorithms

The single-channel algorithm is focused on low-frequency energy since strong energy regions at very low frequencies are commonly observed in speech waveforms in the presence of pop noise. Following, the evolution of the long-term low-frequency is evaluated in order to detect the presence of pop noise. Although this algorithm showed promising performance for different speakers and microphone conditions, its performance was degraded when the input signal came from loudspeakers.

The double-channel algorithm detects pop noise by using a procedure for subtraction between two channels. The setup requires two microphones, one with a pop noise filter and one without, as shown in Figure 2. Let $F_x(b,w)$ and $F_p(b,w)$ be the short-time Fourier transforms (STFT) of the filtered speech and non-filtered speech respectively, where b and w denote the frame index and frequency. Under the assumption that only $F_p(b,w)$ includes pop noise, a differential waveform is estimated by subtracting the ordinary speech component from $F_p(b,w)$ by using $F_x(b,w)$;

$$D(b,\omega) = F_p(b,\omega) - C(\omega)F_x(b,\omega), \tag{1}$$

where $C(\omega)$ represents a compensation filter for compensating between the frequency characteristics of the two channels. An estimate of $C(\omega)$ used to minimize $\sum_{b,\omega}|D(b,\omega)|^2$ can be represented as

$$C(\omega) = \frac{\sum_b F_p(b,w)F_x(b,w)^*}{\sum_b |F_x(b,w)|^2}, \tag{2}$$

where * denotes complex conjugate. The inverse STFT of the subtracted signal $D(b,\omega)$ is assumed to contain information related to pop noise rather than channel conditions or background noise. More precisely, an amplitude-based decision is taken to

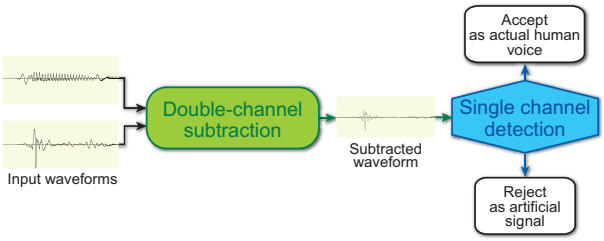

Figure 3: Flow of tandem single-double channel pop noise detection algorithm

characterize the presence of pop noise. Although its performance was not better on average than that of the single-channel algorithm, the double-channel algorithm performed better for different signal conditions without being significantly affected by other sources of noise.

3.2. Tandem single-double channel pop noise detection algorithm

Since the single- and double-channel algorithms both exhibited different benefits and drawbacks, we saw an advantage to integrating them into a single detector to better distinguish an actual human voice from a spoofing attack. We thus created a tandem single-double channel detection algorithm, as shown in Fig. 3. The input waveform is first processed using double-channel subtraction. The subtracted signal is then processed using single-channel detection. This strategy should result in better detection of irregular modulations in the subtracted signal than in the original waveform. Such modulations indicate the presence of pop noise. As a result, detection performance should be better than with the two individual algorithms.

4. Experiments

4.1. Database

Since the proposed framework focuses on speech signals that include pop noise, a database of speech signals including instances of this phenomenon is required. The NIST Speaker Recognition Evaluation (SRE) database [31] is widely used as material for evaluating ASV systems. However, it is not appropriate for our purposes since it provides conversational telephone speech with limited content of pop noise signals. Therefore, we created a new database containing pop noise signals [25]. To evaluate performance, we used three types of microphones:

Voice Microphone with a voice recorder (SONY ECM-DM5P)

Camcorder Compatible microphone with camcorder (SONY ECM-XYST1M)

Headset Microphone with a headset (SHURE SM10A-CN)

Two microphones of each type were used (one with a pop filter), creating a configuration of six microphones channels.

We recorded a 17 female speakers (in Japanese). Each speaker read out 100 sentences in total. Half of the sentences were common to all the speakers and the other half were randomly selected from Japanese Newspaper Article Sentences (JNAS) [32] with a set of randomly selected sentences for each speaker. The 50 common sentences were chosen on the basis of phonetic coverage. We also pre-selected relatively short sentences from the JNAS corpus before the random selection of the rest of the remaining 50 sentences.

Table 1: Equal error rates of VLD algorithms

Microphone	Camcorder	Voice	Headset
Single ch.	4.73%	36.06%	3.95%
Double ch.	29.11%	45.52%	5.88%
Tandem	**0.95%**	26.61%	**2.35%**

4.2. Experimental conditions

We used 30 randomly selected utterances for each microphone without the pop filter for each speaker as live samples of test data. The spoofing materials were constructed on the basis of the statistical parametric speech synthesis framework described by [5]. The speaker adaptation techniques in this framework enable the generation of a synthetic voice using as little as a few minutes of recorded speech from the target speaker [33]. The speaker adaptation algorithm used was structural variational Bayesian linear regression [34]. We used 50 common sentences recorded with the headset microphone with the pop filter to mimic the speaker adaptation of speech synthesis systems (because a pop filter is always used for normal recordings of speech synthesis data). Using the speech synthesizers of individual target speakers, we synthesized artificial speech signals for spoofing. The texts used for speech synthesis were the randomly selected utterances of each speaker. The spoofing materials were then played through a loudspeaker (BOSE 111AD). For the ASV system, we used the standard GMM-UBM-based speaker verification method [35], and the speaker-dependent models of individual speakers in the ASV system were constructed using the 50 common and 20 randomly selected sentences of each speaker recorded with the headset microphone with a pop filter. Here we focus on the effectiveness of the VLD module and not on using a state-of-the-art ASV system. The number of mixtures was set to 2048, and the UBM was trained using about 23,000 utterances from the JNAS database [32], which is the standard speech database for automatic speech and speaker recognition in Japan. For the STFT analysis, the Hamming window was selected as the window function; the window width and the window shift were set to 4096 and 2048 points.

4.3. Experimental results

Table 1 shows the equal error rate (EER) of each VLD algorithm for the three kinds of microphones. When the false positive rate (the percentage of misclassified live voice events) is equal to false negative rate (percentage of misclassified artificial voice events), the common values is the EER. Note that the distance between the speaker's mouth and the microphone varied with the kind of microphone. With conventional methods, the headset microphone generally performs better than camcorder and voice recorder microphones because the mouth is closer to the microphone. With our tandem algorithm, the camcorder microphone resulted in the lowest EER, and the tandem algorithm had the best performance under all microphone conditions. Comparison of the tandem algorithm with the single-channel algorithm show that the EER with the camcorder was reduced from 4.73% to 0.95%. Since the camcorder is the most sensitive microphone with the best noise suppression, noises with a differ-

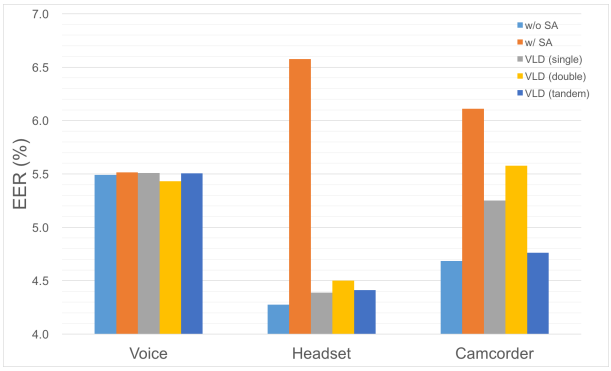

Figure 4: Equal error rates for ASV including spoofing attacks (SA)

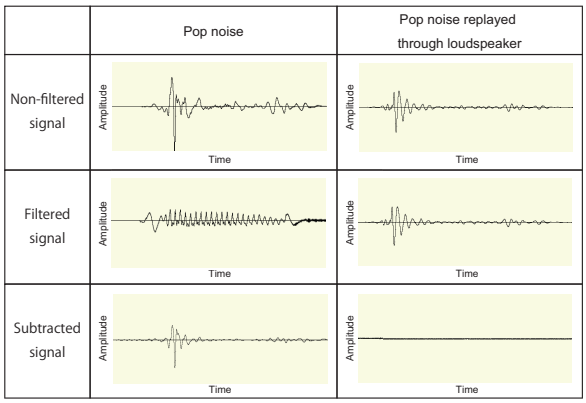

Figure 5: Comparison of waveforms with pop noise before (left) and after replay (right). The two top ones on each side correspond to the two filters shown in Fig 2; the bottom ones are the result of subtraction using the double-channel algorithm.

ent nature may be captured with the single-channel algorithm. In contrast, the tandem algorithm appears to be able to substract some of the common noises denoting the camcorder recordings as the one showing best conditions for pop noise detection.

Figure 4 illustrates the EERs for ASV, including spoofing attacks. The "VLD" denotes the integration of a VLD module into the ASV system. Three different VLD implementations are compared. As expected, spoofing attacks degraded with ASV performance. Since the spoofing attacks were made by enrollment speech recorded with a headset microphone, they were weaker with the voice microphone. The EERs with the headset and camcorder microphones were greatly affected by the presence of spoofing attacks. Moreover, the EERs values for all VLD+ASV cases, clearly reduced the vulnerability of the ASV system. Thses results demonstrate the potential of the proposed framework as an anti-spoofing countermeasure based on voice liveness detection.

4.4. Analysis of effectiveness against replay attack

In the experiment described above, the spoofing attack was Hidden Markov model (HMM)-based speech synthesis for a text-independent ASV system. However, a text-dependent ASV system may also suffer replay attacks. In this case, when pop noise is present in a recordings made by an impostor as enrollment material, it may also appear on the spoofing attacks. As shown by the two top-right waveforms in Fig 5 with the single-channel algorithm, pop noise replayed through a loudspeaker could sometimes be detected while it disappeared in the subtracted waveform with the double-channel algorithm (bottom-right waveform). This implies that the tandem algorithm is effective for both text-dependent and text-independent ASV systems and is thus an effective solution.

5. Conclusion

Identification of "liveness" information in the input speech is needed to protect against spoofing attacks on ASV systems. Two algorithms based on single- and double-channel approaches for capturing this information in terms of the detection of pop noise were previously presented. The tandem single-double channel algorithm presented here improves detection performance. It detects pop noise more accurately and thus improves the discrimination of live voice signals and artificial ones. With this approach, the voice liveness detection performance was significantly improved. Future work includes conducting trials using a larger database and extending the VLD algorithms to strategies based on time-domain features. The robustness of the tandem approach should also be verified on a larger database to better establish its performance under realistic application conditions.

6. Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research (B), 26280066.

7. References

[1] A.K. Jain, P. Flynn, and A.A Ross, "Handbook of biometrics," 2007.

[2] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Audio- and Video-Based Biometric Person Authentication*, vol. 2091 of *Lecture Notes in Computer Science*, pp. 348–353. Springer Berlin Heidelberg, 2001.

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

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

[5] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[6] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing. ICASSP-96. Conference Proceedings., IEEE International Conference on*, May 1996, vol. 1, pp. 373–376 vol. 1.

[7] Y. Stylianou, "Voice transformation: A survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 3585–3588.

[8] Nick Evans, Tommy Kinnunen, and Junichi Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Interspeech*, 2013, pp. 925–929.

[9] Nicholas W D Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre, and Phillip De Leon, *Speaker recognition anti-spoofing*, Book Chapter in "Handbook of Biometric Anti-spoofing", Springer, S. Marcel, S. Li and M. Nixon, Eds., 2014, June 2014.

[10] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[11] Lian-Wu Chen, Wu Guo, and Li-Rong Dai, "Speaker verification against synthetic speech," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, Nov 2010, pp. 309–312.

[12] Zhi-Zheng Wu, Chng Eng Siong, and Haizhou Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition.," in *INTERSPEECH*, 2012.

[13] Marcos Faundez-Zanuy, Martin Hagmller, and Gernot Kubin, "Speaker verification security improvement by means of speech watermarking," *Speech Communication*, vol. 48, no. 12, pp. 1608 – 1619, 2006.

[14] M.A Nematollahi, S.A.R. Al-Haddad, Shyamala Doraisamy, and M. Ranjbari, "Digital speech watermarking for anti-spoofing attack in speaker recognition," in *Region 10 Symposium, 2014 IEEE*, April 2014, pp. 476–479.

[15] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the i -vector space," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 4, pp. 821–832, April 2015.

[16] R.D. McClanahan, B. Stewart, and P.L. De Leon, "Performance of i-vector speaker verification and the detection of synthetic speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3779–3783.

[17] J. Galka, M. Grzywacz, and R. Samborski, "Playback attack detection for text-dependent speaker verification over telephone channels," *Speech Communication*, vol. 67, pp. 143 – 153, 2015.

[18] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, and M. Sahidullah A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech*, 2015, pp. 2037–2041.

[19] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG - Proceedings of the International Conference of the*, Sept 2012, pp. 1–7.

[20] D. Yambay, J.S. Doyle, K.W. Bowyer, A. Czajka, and S. Schuckers, "Livdet-iris 2013 - iris liveness detection competition 2013," in *Biometrics (IJCB), 2014 IEEE International Joint Conference on*, Sept 2014, pp. 1–8.

[21] N. Evans, S.Z. Li, S. Marcel, and A. Ross, "Guest editorial special issue on biometric spoofing and countermeasures," *Information Forensics and Security, IEEE Transactions on*, vol. 10, no. 4, pp. 699–702, April 2015.

[22] Bori Toth, "Liveness detection: Iris," in *Encyclopedia of Biometrics*, pp. 931–938. Springer US, 2009.

[23] StephanieA.C. Schuckers, "Liveness detection: Fingerprint," in *Encyclopedia of Biometrics*, pp. 924–931. Springer US, 2009.

[24] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision ECCV 2010*, vol. 6316 of *Lecture Notes in Computer Science*, pp. 504–517. Springer Berlin Heidelberg, 2010.

[25] S. Shiota, F. Villavicencio, 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," in *Interspeech*, 2015, pp. 239–243.

[26] G.W. Elko, Jens Meyer, Steven Backer, and J. Peissig, "Electronic pop protection for microphones," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, Oct 2007, pp. 46–49.

[27] Yimin Hsu, "Spectrum analysis of base-line-popping noise in mr heads," *Magnetics, IEEE Transactions on*, vol. 31, no. 6, pp. 2636–2638, Nov 1995.

[28] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Acoustics, Speech, and Signal Processing. ICASSP-93., IEEE International Conference on*, April 1993, vol. 2, pp. 391–394 vol.2.

[29] D.P. Delacretaz and J. Hennebert, "Text-prompted speaker verification experiments with phoneme specific mlps," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, May 1998, vol. 2, pp. 777–780 vol.2.

[30] J. Sanchez, I. Saratxaga, I. Hernez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[31] "NIST Speaker Recognition Evaluation (SRE)," http://www.itl.nist.gov/iad/mig/tests/spk/.

[32] K. Ito, 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," *the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[33] Junichi Yamagishi and Takao Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[34] Shinji Watanabe, Atsushi Nakamura, and Biing-Hwang(Fred) Juang, "Structural bayesian linear regression for hidden markov models," *Journal of Signal Processing Systems*, vol. 74, no. 3, pp. 341–358, 2014.

[35] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verificaion using adapted gaussian mixture models," *Degital Signal Orocess*, vol. 10, no. 1, pp. 19–41, 2000.