Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

Voice liveness detection using phoneme-based pop-noise detector for speaker verification

Shihono Mochizuki, Sayaka Shiota, Hitoshi Kiya

Department of Information and Communication Systems,
Tokyo Metropolitan University, Tokyo, Japan

Abstract

This paper proposes a phoneme-based pop-noise (PN) detection algorithm for voice liveness detection (VLD) and automatic speaker verification systems. Recently, a lot of countermeasures against spoofing attacks (e.g., replay, speech synthesis) have been reported for speaker verification systems. A principle mechanism of almost all spoofing attacks is to replay recorded speeches via a loudspeaker. Therefore, one of the effective solutions against spoofing attacks is to determine whether an input speech is a genuine voice or a replayed one, and this is a framework of VLD. To realize the VLD framework, PN detection methods have been proposed. Since PN is a common distortion that occurs when speaker's breath reaches the inside of a microphone, the conventional PN detection methods simply capture PN periods during the input speech. However, the performances of the PN detection methods depend on microphone types and phrases. It may lead to vulnerability of the conventional PN detection methods. This paper proposes a novel PN detection method, focused on specific characteristics of phonemes related to the PN phenomenon. The experimental results show that the proposed method provides a higher performance than conventional PN detection methods.

1. Introduction

Biometric authentication plays an important role in reliable management systems [1]. Automatic speaker verification (ASV) is an easy-to-use biometric authentication using speech. State-of-the-art ASV systems, as i-vector-based [2–4] and probabilistic linear discriminant analysis (PLDA)-based ones [5–7], have achieved reliable performances and are expected to be in practical use. Meanwhile, performances of speech synthesis algorithms such as text-to-speech (TTS) systems [8–10] and voice conversion systems [11, 12] have been significantly improved. By using these synthesis algorithms, arbitrary speeches of a target speaker can be easily generated, and the generated speeches can behave like a key to the ASV systems. Therefore, spoofing attacks with synthesis algorithms and recorded speeches have become a serious problem for ASV systems [13–15].

ASV spoofing and countermeasures challenge (ASVspoof 2015) was held as a research event regarding the performance of ASV systems against spoofing attacks [16]. ASVspoof 2015 focused on countermeasures for synthesis-based spoofing attacks. While these attacks require some technical knowledge, for imposters recording and replaying can be performed easily via consumer devices. Therefore, at the second challenge (ASVspoof 2017 [17]), replay attacks were used as spoofing attacks. In biometrics theory applications and systems (BTAS) 2016, a speaker anti-spoofing challenge was held to detect replay attacks by introducing an AVspoof database [18]. A principle approach of almost all spoofing attacks is to replay recorded speeches via a loudspeaker. Thus, an effective solution against spoofing attacks is to determine whether the input speech is a genuine human voice or a replayed one, and this is a framework of voice liveness detection (VLD) [19]. To realize the VLD framework, pop-noise (PN) detection methods have been reported [20]. Since PN is a common distortion that occurs when a speaker's breath reaches the inside of a microphone [21], it is expected that PN phenomenon can be a distinguishable characteristic of a genuine speech from a replayed one. It has also been reported that the PN detection methods obtained a high accuracy in detecting spoofing attacks [20]. However, the performances of PN detection methods depend on microphone types and uttered content. The conventional PN detection method simply captures the PN periods during the input speech. However, the PN phenomenon can also be caused by an imposter's breath. It may lead to the vulnerability of conventional PN detection methods.

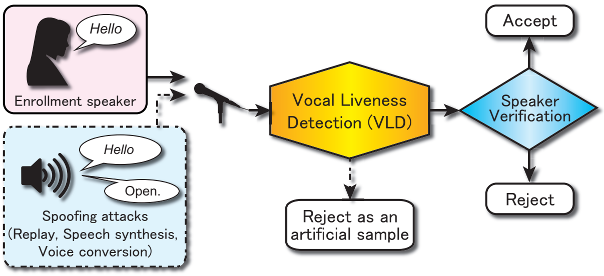

Figure 1: Diagram of VLD and ASV systems

This paper proposes a novel PN detection method, focused on specific characteristics of phonemes related to the PN phenomenon. Considering the process of pronunciation, there are specific phoneme groups: one requires much breathing and another group hardly requires any breathing. In the proposed algorithm, phonemes in detected PN periods are compared with both phoneme groups. When the phonemes belong to the adequate group, it can be used as evidence of liveness. Additionally, PN-balanced sentences are designed as prompt sentences for experiments. It is also discussed that selection methods of suitable phoneme groups for the proposed method. The experimental results demonstrate that the proposed algorithm provides a higher performance than conventional PN detection methods.

In section 2, we briefly describe the voice liveness detection framework. Our proposed algorithm is presented in section 3. Section 4 illustrates experimental results. Finally, concluding and future work are given in section 5.

10.21437/Odyssey.2018-33

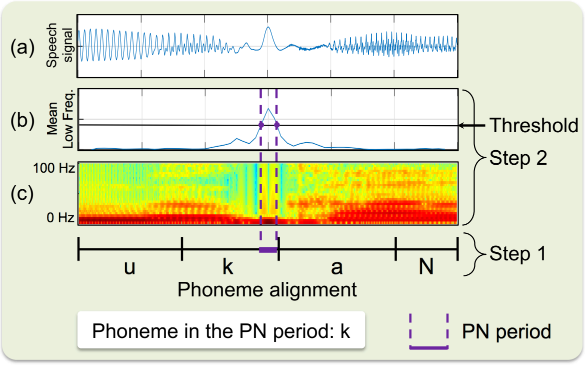

Figure 2: Phoneme selection in PN period

2. Single-channel pop-noise (PN) detection algorithm and phoneme selection in PN periods

Figure 1 shows a diagram of a VLD and an ASV system. The VLD system is designed to reject all speeches which show no evidence of liveness. Only accepted speeches from the VLD system are evaluated in the ASV system.

To realize the VLD framework, a single-channel PN detection method has been reported [19]. Pop-noise (PN) is commonly known as a phenomenon that occurs when a speaker's breath reaches the inside of a microphone [21, 22]. Thus, it is expected that PN phenomenon can be a distinguishable characteristic separating a genuine speech from a replayed one. Considering the PN phenomenon, power spectrum follows a track of an irregular, strong, and instant peak, and this characteristic is remarkable at a very low-frequency bin (Fig. 2(b)). Therefore, the single-channel detection method captures PN periods where a trajectory is over a threshold (Fig. 2(c)). It has also been reported that the PN detection methods obtained a high accuracy to detecting in spoofing attacks [20]. Additionally, since the single-channel PN detection method can be performed with one microphone, this method is easily feasible and connection to ASV systems.

3. Phoneme-based PN detection

3.1. Relation between PN phenomenon and phonemes

The conventional PN detection methods simply capture PN periods during the input speech. However, performances of the PN detection methods depend on microphone types and uttered content. It may lead to the vulnerability of conventional PN detection methods.

This paper proposes a novel PN detection algorithm, focused on specific characteristics of phonemes related to the PN phenomenon. Considering a process of pronunciation, there are specific phoneme groups: one requires much breathing and another group hardly requires any breathing. In the proposed algorithm, phonemes in the detected PN periods are compared with both phoneme groups. When the phonemes belong to the adequate group, it can be used as evidence of liveness.

3.2. Phoneme selection in PN periods

The procedure of phoneme selection in the detected PN periods is shown as below:

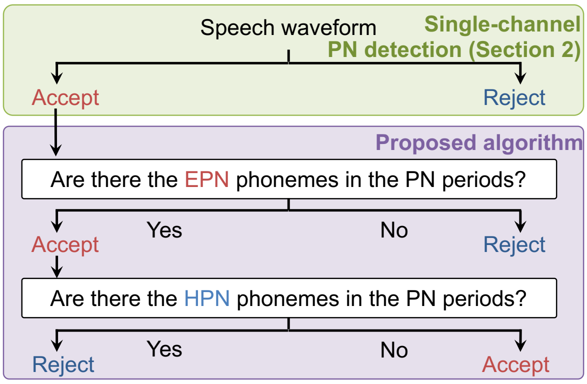
Speech waveform
Single-channel PN detection (Section 2)
Accept
Reject
Proposed algorithm
Are there the EPN phonemes in the PN periods?
Yes
No
Accept
Reject
Are there the HPN phonemes in the PN periods?
Yes
No
Reject
Accept

Figure 3: Flow of proposed algorithm (EPN: easily PN phenomenon caused, HPN: hardly PN phenomenon caused)

Step 1: Estimate phoneme alignment of an input speech by using an automatic speech recognition system.

Step 2: Detect some PN periods with the single-channel PN detection method (Sec. 2).

Step 3: By comparing the detected PN periods with the estimated phoneme alignments, some phonemes are selected (Fig. 2). The selected phonemes with a strict threshold value means that they easily caused PN phenomenon.

Step 4: Some process of Step 3 with a sufficiently low threshold value. In this case, almost phonemes are selected as in the PN periods. Some phonemes, which have not been selected, mean that they hardly caused PN phenomenon.

The selected phonemes in Step 3 are called EPN phonemes, and the phonemes which not been selected in Step 4 are called HPN phonemes.

3.3. Proposed algorithm

Figure 3 illustrates a flow of the phoneme-based PN detection algorithm. At first, PN periods are detected from the input speech by using the single-channel PN detection method. When the speech has some PN periods, the speech is accepted as a genuine speech. When there are no PN periods in the speech, the speech is rejected as a replay attack. In the second step, the proposed phoneme-based PN detection is performed. By using the procedure of Sec. 3.2, some phonemes are picked up and compared with the EPN phoneme group. If there are EPN phonemes in the PN periods, the speech is accepted as a genuine speech. However, since the PN phenomenon is sometimes caused by accidental wind, it is also important that the HPN phonemes do not appear in the PN periods. When a human pronounces HPN phonemes, it is hard to generate the PN phenomenon. Therefore, the case that there are HPN phonemes in PN periods is unnatural for genuine speech, and the speech is rejected as a replay attack. By using this algorithm, replay attacks, accepted in the conventional PN detection method, are rejected, and the reliability of the VLD system improves.

3.4. Selection schemes of EPN and HPN phonemes

To improve the performance of the proposed method, selection method of effective phoneme group is also considered. EPN and HPN phoneme groups are dependent on speaking style for

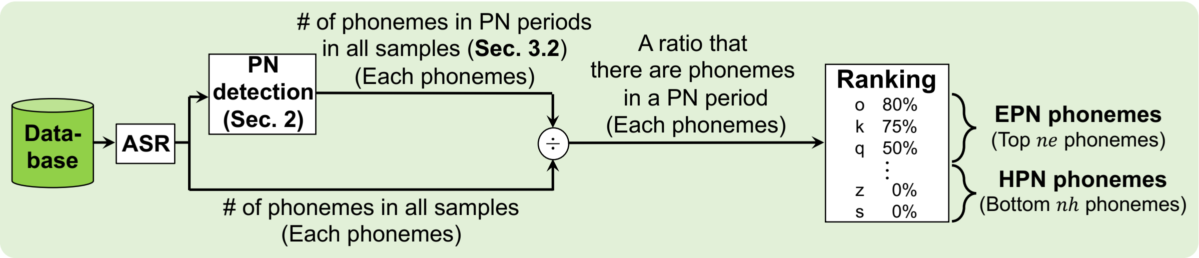
Data-base
ASR
PN detection (Sec. 2)
# of phonemes in PN periods in all samples (Sec. 3.2) (Each phonemes)
# of phonemes in all samples (Each phonemes)
÷
A ratio that there are phonemes in a PN period (Each phonemes)
Ranking
o 80%
k 75%
q 50%
z 0%
s 0%
EPN phonemes (Top ne phonemes)
HPN phonemes (Bottom nh phonemes)

Figure 4: Diagram of selection scheme of common phoneme group

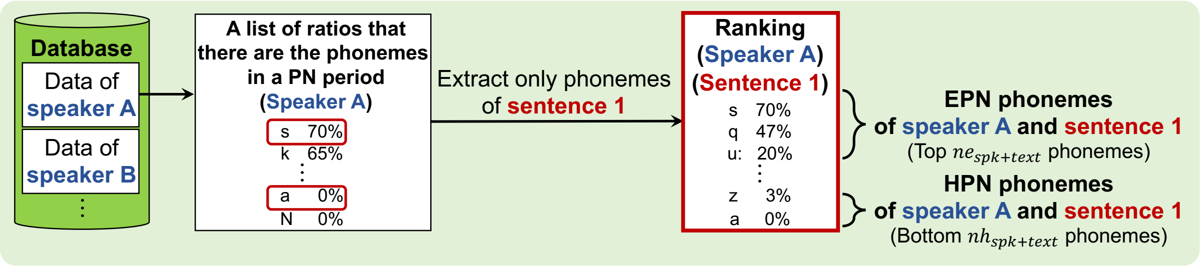
Database
Data of speaker A
Data of speaker B
A list of ratios that there are the phonemes in a PN period (Speaker A)
s 70%
k 65%
a 0%
N 0%
Extract only phonemes of sentence 1
Ranking (Speaker A) (Sentence 1)
s 70%
q 47%
u: 20%
z 3%
a 0%
EPN phonemes of speaker A and sentence 1 (Top $ne_{spk+text}$ phonemes)
HPN phonemes of speaker A and sentence 1 (Bottom $nh_{spk+text}$ phonemes)

Figure 5: Diagram of selection scheme of speaker- and text-dependent phoneme group

Table 1: Recording conditions

VLD database	
Microphone	SONY ECM-XYST1M
Loudspeaker	ELECOM LBT-SPP300
# of speakers	17 female speakers
Sampling rate	48 kHz

VLD2 database	
Microphone	AKG P170
Loudspeaker	ELECOM LBT-SPP300
# of speakers	15 female speakers
Sampling rate	48 kHz

Table 2: Test conditions

Database	VLD2 database
Sampling rate	48 kHz
# of speakers	15 speakers
# of genuine speeches	10 sentences / speaker
# of replay attacks	10 sentences / speaker

Table 3: Experimental conditions of single-channel PN detection method

Sampling rate	48 kHz
Bit rate	24 bit
Frequency bin	[0, 50] Hz
Window length	20 msec
Window shift	4 msec

Table 4: Dataset to select EPN and HPN phonemes of each phoneme group

Common phoneme group	
Database	VLD database
Sampling rate	48 kHz
# of speakers	17 speakers
# of sentences	30 sentences / speaker
# of genuine speeches	510 samples

Speaker- and text-dependent phoneme group	
Database	VLD database
Sampling rate	48 kHz
# of speakers	8 speakers
# of sentences	120 sentences / speaker
# of genuine speeches	960 samples

each speaker. Thus, it is expected that the performance of the proposed method is improved by using speaker-dependent EPN and HPN phoneme groups. Meanwhile, text-dependent speaker verification systems are used for smartphone security and banking systems in recent years. Therefore, EPN and HPN phoneme groups are also dependent on the text-dependent condition. For these reasons, it is expected that the performance of the proposed method is improved by using speaker- and text-dependent phoneme groups.

4. Evaluation experiments

To evaluate the effectiveness of the proposed method, VLD experiments were performed with using Japanese database and English one.

4.1. Japanese database experiment

4.1.1. Experimental conditions

VLD database [19] and VLD2 database [23] were used for VLD experiments. VLD and VLD2 database are Japanese database for evaluation of the PN detection methods. Since the genuine utterances in these databases were recorded by a microphone without any pop filter, the PN phenomenon frequently occurred. These database includes genuine speeches and replay attacks. The recording conditions are shown in Tab. 1. All samples were recorded in a sound-proof room. The distance

Table 5: Common phoneme group and examples of speaker- and text-dependent phoneme group

Common phoneme group				
All speakers	All sentences	EPN	t, ky,hy,b,s,sh,k,o:,e:,u:,o	
		HPN	ry,i,m	

Speaker- and text-dependent phoneme group (example)				
		Phoneme group	Same phonemes as in the common phoneme group	Different phonemes from the common phoneme group
Speaker 1	Sentence 1	EPN	o,o;,sh	a,d,e,j,m,n,N,py
		HPN	-	e:
	Sentence 2	EPN	e:,k,o,o:,s,t	a,i,n,N,py
		HPN	-	d

between the microphone and each speaker's mouth or the loudspeaker was about 7 cm. The phrases are phoneme balanced sentences base on Japanese news article sentences. The test sets are shown in Tab. 2, and the conditions of the single-channel PN detection method are shown in Tab. 3. The threshold of the single-channel PN detection method was set at the point where genuine speeches of 90% were accepted. An open-source large vocabulary speech recognition engine Julius [24] and its dictation kit with DNN-HMM method was used to perform speech recognition and to obtain monophone alignments. To evaluate the effectiveness of using speaker- and text-dependent phoneme group, two types of phoneme groups were prepared. One type is called common phoneme group, which selected from VLD database (Fig. 4). The other one is called speaker- and text-dependent phoneme group, which selected from each speaker and phrase (Fig. 5). Tab. 4 shows a dataset in order to select these phoneme groups. Both number of the common EPN phonemes ne and of the speaker- and text-dependent EPN phonemes $ne_{spk+text}$ were set to 11. Both the number of the common HPN phonemes nh and a speaker- and text-dependent HPN phoneme $nh_{spk+text}$ were set to 3 and 1 respectively. The selected phoneme groups are shown in Tab. 5. These numbers were decided from preliminary experiments. At first, it considered that the well-known plosive phonemes are appropriate for the common phoneme group. However, from our preliminary investigation, it had clarified that only plosive phonemes are not necessarily as the EPN phonemes. Therefore, the actual EPN and HPN phonemes were selected from the extraction procedure which written in section 3.2. For the evaluation measure of the VLD frameworks, false rejection rate (FRR) and false acceptance rate (FAR) were used.

$$\mathrm{FRR} = \frac{\text{\# of rejected genuine utterances}}{\text{\# of all genuine utterances}}, \tag{1}$$

$$\mathrm{FAR} = \frac{\text{\# of accepted spoofing attacks}}{\text{\# of all spoofing attacks}}. \tag{2}$$

The proposed VLD methods require to set the threshold of the single-channel PN detection at first, then the EPN and HPN phonemes are evaluated with the detected PN periods. Thus, since the evaluation measure depends on the threshold, the equal error rate was not used for the experiments.

Figure 6 shows a flow of each compared algorithm. The details are shown as below:

(A) PN detection (Conventional method):
Identify an input signal as a genuine speech or a spoofing attack by the single-channel PN detection.

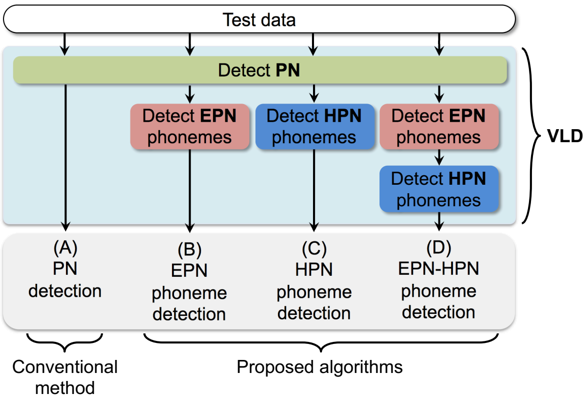

Figure 6: Flow of each algorithm

(B) EPN phoneme detection:
After the PN detection process, the detected phonemes from the PN periods are compared with the EPN phonemes.

(C) HPN phoneme detection:
After the PN detection process, the detected phonemes from the PN periods are compared with the HPN phonemes.

(D) EPN-HPN phoneme detection:
After the EPN phoneme detection (B) process, the detected PN periods of accepted utterances are additionally compared with HPN phonemes (Fig. 3).

4.1.2. Experimental results

Figure 7 and 8 illustrate FRRs and FARs of each algorithms with the common phoneme group and the speaker- and text-dependent phoneme group respectively. From FARs of the common phoneme group (Fig. 7), FARs of (B) and (D) are lower than that of (A). However, in Fig. 8, FRRs of (B) and (D) are higher than that of (A). These results illustrates that the proposed method can reject spoofing attacks more precisely. However, since FRRs of the phoneme-based method were higher than the conventional method, it may cause usability problem. On the other hand, the FRRs of the speaker- and text-dependent phoneme group are lower than those of the common phoneme group, even though FARs are almost same as the common phoneme group ones. These results indicate that the performance of the proposed method is improved by using speaker-

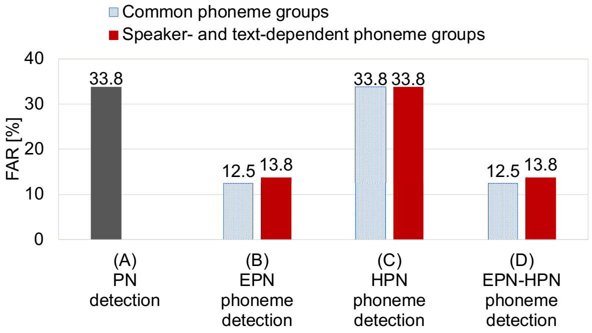

Figure 7: FARs with common phoneme group and speaker- and text-dependent phoneme group

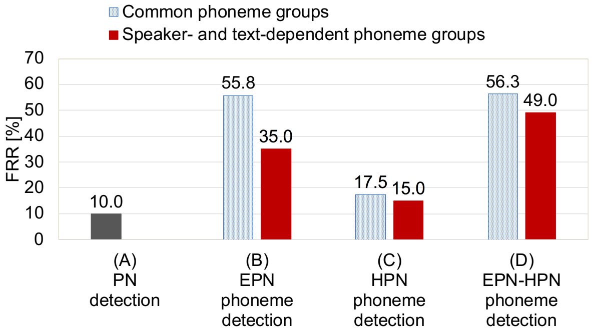

Figure 8: FRRs with common phoneme group and speaker- and text-dependent phoneme group

Table 6: Test conditions

Database	AVspoof database (Test set)
# of sessions	1
Sampling rate	16 kHz
# of speakers	18 speakers
# of genuine speeches	5 sentences / speaker
# of replay attacks	5 sentences / speaker

Table 7: Experimental conditions of single-channel PN detection method

Sampling rate	16 kHz
Bit rate	16 bit
Frequency bin	[0, 65] Hz
Window length	15.4 msec
Window shift	7.7 msec

and text-dependent phoneme group. However. there is still remained the vulnerability against spoofing attacks.

4.2. English database experiment

4.2.1. Experimental conditions

AVspoof database [25], which is one of the English database for spoofing attacks, was used. The AVspoof database provides stable, non-biased spoofing attacks in order to test anti-spoofing algorithms for ASV systems. The AVspoof database includes genuine speeches and three types of spoofing attacks. The spoofing attacks consist of replay attacks, speech syntheses and voice conversions. These audio data are recorded by a high-quality microphone (AT2020USB+) and smartphones (Samsung Galaxy S4, Iphone 3GS). The phrases are divided into three parts, reading part, pass-phrase part and free speech part. In this experiment, genuine speeches and replay attacks of pass-phrase part recorded by a high-quality microphone were used. The test conditions are shown in Tab. 6. The conditions of the single-channel PN detection method are shown in Tab. 7. The threshold of the single-channel PN detection method was set at the point where genuine speeches of 90% were accepted. To estimate monophone alignment, Julius and its dictation kit with GMM-HMM. The acoustic model which was estimated from speech corpora of book excerpts and broadcast news was used [26]. The language model which was suitable for the pass-phrases was used. Both number of the common EPN phonemes ne and of the speaker- and text-dependent EPN phonemes $ne_{spk+text}$ were set to 11. Both the number of a common HPN phoneme nh and a speaker- and text-dependent HPN phoneme $nh_{spk+text}$ were set to 1. These numbers were decided from preliminary experiments. The datasets for selection of EPN and HPN phonemes are shown in Tab. 8. The examples of common phoneme group and the speaker- and text-dependent phoneme group are shown in Tab. 9. Experimental flow and the evaluation measures were same to Sec. 4.1.

Table 8: Dataset to select EPN and HPN phonemes of each phoneme group

Common phoneme group	
Database	AVspoof database (Training set)
# of sessions	4
Sampling rate	16 kHz
# of speakers	14 speakers
# of sentences	5 sentences / speaker
# of genuine speeches	280 samples
Speaker- and text-dependent phoneme group	
Database	AVspoof database (Test set)
# of sessions	3
Sampling rate	16 kHz
# of speakers	18 speakers
# of sentences	5 sentences / speaker
# of genuine speeches	270 samples

4.2.2. Experimental results

Figure 9 and 10 illustrates FRRs and FARs with the common phoneme group and the speaker- and text-dependent phoneme group. Similar tendencies to the Japanese experiment in sec.4.1 can be observed. It indicates that the proposed method works on the other language.

The experimental results of the proposed method in this section are not good since the AVspoof database is not recorded for PN detection. Nevertheless, the results of PN detection are good. Because the test data were specified by recording device in order to be easy to do PN detection. However, if imposters intentionally cause the PN phenomenon during spoofing attack replays, it is concerned that FAR of PN detection increase. To reject such attacks against the PN detection, considering phoneme information is necessary. From the experimental results, it appears that only the EPN phoneme detection is need. However,

Table 9: Common phoneme group and examples of speaker- and text-dependent phoneme group

Common phoneme group				
All speakers	All sentences	EPN	aw,dh,eh,er,f,hh,k,m,ow,s,t	
		HPN	aa	

Speaker- and text-dependent phoneme group (example)				
		Phoneme group	Same phonemes as in the common phoneme group	Different phonemes from the common phoneme group
Speaker 1	Sentence 1	EPN	er,s,t	ah,ih,iy,n,p,r,sh,uw
		HPN	-	w
	Sentence 2	EPN	aw,hh,m,s	aa,ih,iy,n,p,sh,uw
		HPN	-	dh

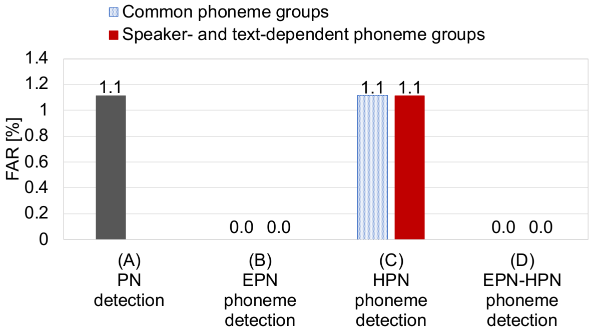

Figure 9: FARs with common phonemes and speaker- and text-dependent phonemes

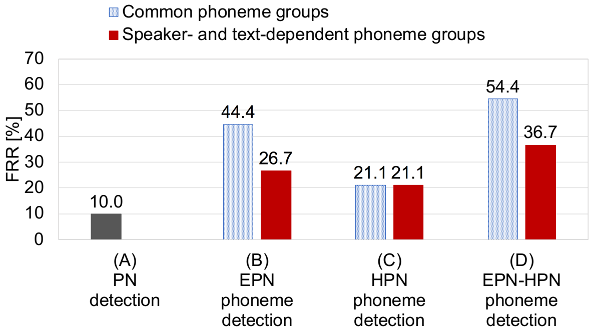

Figure 10: FRRs with common phonemes and speaker- and text-dependent phonemes

the HPN phoneme detection is also importance as a countermeasure of attacks agains the PN detection.

5. Conclusions

This paper proposed the phoneme-based PN detection algorithm focused on the specific characteristics of phonemes related to the PN phenomenon. In the proposed algorithm, phonemes in the detected PN periods were compared with the EPN and the HPN phonemes to classify genuine speeches from spoofing attacks. The experimental results demonstrated that the proposed algorithm with speaker- and text-dependent phoneme group provided higher performance than the conventional PN detection methods. Moreover, evaluation using two databases indicated that the proposed algorithm is robust for differences of database and language.

Future work includes a creating a database which can be used against both VLD systems and ASV systems and conducting trials of integration between the proposed method and the ASV system is performed. In the experiments, the spoofing attacks were simply replayed via a loudspeaker. One type of aggressive spoofing attack is when imposters intentionally cause the PN phenomenon during spoofing attack replays. Therefore, the proposed method will be performed under that situation. Additionally, a present system of the proposed method is rule-based. Hence, statistic modeling approach of the proposed method is also planning to pursue.

6. Acknowledgements

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

7. References

[1] A. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, Springer Science & Business Media, 2007.

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

[3] H. Zeinali, H. Sameti, L. Burget, J. Černockỳ, N. Maghsoodi, and P. Matějka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge," *in Proc. INTERSPEECH*, pp. 440–444, 2016.

[4] I. Yang, H. Heo, S. Yoon, and H. Yu, "Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using deep neural network," *in Proc. ICASSP*, pp. 5490–5494, 2017.

[5] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *in Proc. ICCV*, pp. 1–8, 2007.

[6] A. Kanagasundaram, D. Dean, S. Sridharan, C. Fookes, and I. Himawan, "Short utterance variance modelling and utterance partitioning for PLDA speaker verification," *in Proc. INTERSPEECH*, pp. 1835–1838, 2016.

[7] S. Madikeri, M. Ferras, P. Motlicek, and S. Dey, "Intra-class covariance adaptation in PLDA back-ends for speaker verification," Tech. Rep., Idiap, 2017.

[8] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *in Proc. ICASSP*, vol. 1, pp. 373–376, 1996.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[10] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[11] Y. Stylianou, "Voice transformation: a survey," *in Proc. ICASSP*, pp. 3585–3588, 2009.

[12] Z. Wen, K. Li, J. Tao, and C. Lee, "Deep neural network based voice conversion with a large synthesized parallel corpus," *in Proc. Signal and Information Processing Association Annual Summit and Conference*, pp. 1–5, 2016.

[13] 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.

[14] N. W. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," *in Proc. INTERSPEECH*, pp. 925–929, 2013.

[15] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of Biometric Anti-Spoofing*, Springer, 2014.

[16] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," *in Proc. INTERSPEECH*, pp. 2037–2041, 2015.

[17] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, "ASVspoof: the automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[18] P. Korshunov, S. Marcel, H. Muckenhirn, A. R. Gonçalves, A. G. S. Mello, R. P. Velloso Violato, F. O. Simões, M. U. Neto, M. de Assis Angeloni, J. A. Stuchi, et al., "Overview of BTAS 2016 speaker anti-spoofing competition," *in Proc. BTAS*, pp. 1–6, 2016.

[19] 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 Proc. INTERSPEECH*, pp. 239–243, 2015.

[20] S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, "Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector," *in Proc. Odyssey 2016*, pp. 259–263, 2016.

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

[22] Y. Hsu, "Spectrum analysis of base-line-popping noise in MR heads," *IEEE transactions on magnetics*, vol. 31, no. 6, pp. 2636–2638, 1995.

[23] S. Mochizuki, S. Shiota, and H. Kiya, "Voice liveness detection based on phoneme information-based pop-noise detector (japanese)," *IEICE Trans. on information and systems*, vol. J101–D, no. 3, 2018.

[24] "Julius, Large vocabulary continuous speech recognition decoder software," `http://julius.osdn.jp`.

[25] S. K. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," *in Proc. BTAS*, pp. 1–6, 2015.

[26] "VoxForge," `http://www.voxforge.org`.