PRIVACY-PRESERVING SOUND TO DEGRADE AUTOMATIC SPEAKER VERIFICATION PERFORMANCE

Kei Hashimoto[1], *Junichi Yamagishi*[2,3], *and Isao Echizen*[2]

[1]Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan
[2]National Institute of Informatics, Tokyo, Japan
[3]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, United Kingdom

ABSTRACT

In this paper, a privacy protection method to prevent speaker identification from recorded speech is proposed and evaluated. Although many techniques for preserving various private information included in speech have been proposed, their impacts on human speech communication in physical space are not taken into account. To overcome this problem, this paper proposes privacy-preserving sound as a privacy protection method. The privacy-preserving sound can degrade speaker verification performance without interfering with human speech communication in physical space. To make a first step toward solving this problem, suitable sound characteristics for preserving privacy are evaluated in terms of the speaker verification performance and speech intelligibility. The experimental results show that appropriate sound can efficiently degrade the speaker verification performance without degrading speech intelligibility.

Index Terms— Speaker verification, privacy protection, speech intelligibility

1. INTRODUCTION

With the popularization of portable devices with built-in microphones, such as smartphones and tablets, and advances in spoken language processing technologies, many attractive systems and applications that analyze speech and sound and provide services have been developed, e.g., Google Now, Apple Siri, and Amazon Echo. On the other hand, there is a problem of speech information being recorded with portable devices and then shared on the Internet, e.g., social networking sites, without the person's permission. Private information, e.g., when and where a person was, is easily revealed by analyzing recorded speech and accessing various information added by portable devices, such as global positioning system (GPS) data [1, 2]. Unauthorized information can potentially be revealed by comparing information on several social networking sites. Furthermore, the privacy problem becomes more serious if a person's identity can be obtained from speech information.

Speaker recognition, which is the process of automatically recognizing who is speaking on the basis of individual information included in speech waveforms, has made very significant progress over the past 50 years [3, 4, 5]. This technique enables the use of voices to verify identities and control access to services such as telephone-based banking. Several services using speaker recognition technologies have already been introduced into practical use, e.g., Barkleys bank[1]. It is expected that speaker recognition technologies create new services that make our daily lives more convenient and such services become popular in the future. The application of speaker recognition techniques for crime, however, also makes it possible to reveal important information by identifying a person from voice characteristics because state-of-the-art automatic speaker recognition systems show higher recognition accuracy than human listeners [6, 7]. Therefore, privacy protection techniques that can prevent such identification from speech information are needed.

Recently, multimedia information such as images, video, audio, and text can be easily posted online and shared through social networking services. At the same time, many privacy protection techniques for such information have been proposed. For speech information, Jin *et al.* proposed a speaker de-identification technique using voice transformation to prevent speaker identity disclosure [8, 9]. De-identification for multimedia information has been defined as the process of concealing the identities of individuals captured in a given set of information. To perform speaker de-identification, however, this technique requires some processes after speech recording. Consequently, if speech is recorded by someone else in physical space, the technique cannot be applied to protect private information. To overcome this problem, we propose "privacy-preserving sound" as a privacy protection technique. The purpose of privacy-preserving sound is to degrade the performance of automatic speaker verification systems without any processes after recording. In addition, the proposed technique is intended not to interfere with human speech communication in physical space. Therefore, the proposed technique can protect a speaker's identity without interfering with communication by generating sound that has an insignificant effect on speech intelligibility. To make a first step toward solving this problem, in this paper we investigate what kind of sound works well as privacy-preserving sound, through speaker verification experiments and objective evaluation of speech intelligibility. Experimental results show that appropriate sound can degrade speaker verification performance without degrading speech intelligibility.

The rest of this paper is organized as follows. Sections 2 and 3 describe related work and the idea of privacy-preserving sound, respectively. The experimental conditions and results are given in Section 4. Concluding remarks and future work are presented in Section 5.

2. RELATED WORK

As privacy issues for multimedia information become more acute, the demand for privacy protection techniques is increasing. Speech signals include both non-linguistic private information, such as a speaker's identity, and linguistic private information. A number of privacy protection techniques have been proposed in the speech processing area.

Jin *et al.* proposed a speaker de-identification technique using

[1] https://wealth.barclays.com

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

voice transformation to prevent speaker identity disclosure [8, 9]. The proposed technique allows a speaker's voice to be de-identified in the sense that it still sounds natural and intelligible but does not reveal the speaker's identity. Therefore, this technique enables transmission of the content of a user's spoken requests while successfully protecting his identity. Parthasarathi *et al.* proposed a privacy-preserving audio representation to protect linguistic private information [10, 11]. The proposed representation includes a speaker's individual information but low linguistic information. Therefore, private information such as the lexical contents included in the original audio files can be protected by storing audio files with the proposed representation instead. In [12], Yamamoto *et al.* proposed techniques for protecting individuality in speech signals and linguistic private information included in speech recognition results. Additionally, they proposed a technique for eliminating the speech from audio files. Since the above techniques can protect private information but require post-processing of recorded speech, they are not appropriate for a situation in which speech is recorded by someone else in physical space and posted online without permission.

A sound masking method that is a sound addition approach for eliminating unwanted speech sound was proposed as another privacy-preserving technique [13, 14]. This technique is developed to protect linguistic private information in physical space. It can thus be used to protect information contained in private conversations in open areas, such as banks, pharmacies, medical examination rooms, and offices. Although this technique does not require post-processing, sound environment and speech intelligibility for people in the area may be affected because this technique need to generate and add sounds that can eliminate unwanted speech sound. Therefore, this sound masking method need to be used at appropriate situation and setting.

3. PRIVACY-PRESERVING SOUND

The state-of-the-art automatic speaker recognition systems show higher recognition accuracy than human listeners [6, 7]. In addition, there is a problem that multimedia information is recorded by someone else in physical space and shared on social networking services without permission. If automatic speaker recognition systems is used for crime, important privacy information will be revealed by identifying a person from voice characteristics and serious privacy problems will be enormously increased. Therefore, in this work, we investigate privacy protection techniques that can prevent speaker identification using automatic speaker verification systems.

Some privacy protection techniques which are previously proposed can be utilized to protect speaker identity, e.g., speaker de-identification techniques using voice transformation [8, 9] and elimination of detected speech region [12]. However, it is difficult to apply these techniques for protecting speaker identities in a situation in which speech is recorded by someone else because they require some processes after recording. There are some protection methods that can be used even if speech information is recorded unintentionally, e.g., a noise generation to create low signal-to-noise ratio (SNR) environments and voice transformation using real-time voice changers. The performance of automatic speaker verification systems can be degraded by adding noise to speech or drastically changing voice characteristics. However, people cannot talk comfortably in a low SNR environment due to the degradation of speech intelligibility, and they cannot communicate naturally if people's voice characteristics are changed from their original characteristics. In other words, these techniques for preventing speech identification degrade the quality of human speech communication.

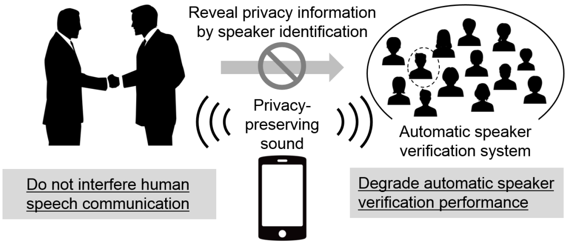

Fig. 1. Overview of privacy-preserving sound

To overcome these problem, we propose "privacy-preserving sound" as a privacy protection technique that can protect a speaker's identity without interfering with human speech communication even if speech is recorded by someone else in physical space. Figure 1 shows an overview of privacy-preserving sound. The purpose of privacy-preserving sound is to degrade the performance of automatic speaker verification systems. In addition, the proposed technique is intended not to interfere with human speech communication. Therefore, by generating a sound that has the property of not affecting speech intelligibility and adding it conversational speech, the proposed method degrade speaker verification performance without interfering with human speech communication. It is expected that sound characteristics, such as SNR, frequencies, and spectral envelopes, show different impacts on speaker verification systems and speech intelligibility. By taking account of such difference, we investigate what kind of sound can be used as privacy-preserving sound.

The proposed method is based on a sound adding approach as like sound masking methods [13, 14]. Table 1 shows properties of privacy-preserving sound and sound masking. As shown in Table 1, the private information preserved by them and target of the sound are different. In standard sound masking methods, the target private information is linguistic information contained in private conversations in open areas. On the other hand, the proposed privacy-preserving sound is used to prevent speaker identity revelation with automatic speaker verification systems. Therefore, different sound properties are needed for these methods and impacts of these methods on speech communication in physical space are different. The proposed method can be used to protect private information in various situations because the sound that does not affect speech intelligibility significantly but has strong impacts on the performance of automatic speaker verification systems is generated for privacy protection.

4. EXPERIMENTS

The goal of our study is to protect privacy by preventing speaker identification with automatic speaker verification systems, without interfering with human speech communication in physical space. To make a first step toward developing privacy-preserving sound, we conducted speaker verification experiments and objective evaluation of speech intelligibility to investigate what kind of sound would work well. In these experiments, various sounds based on white noise were added to test data. In particular, we focused on the SNR and frequency of sound, and investigated their impact on speaker verification performance and speech intelligibility.

Table 1. Property of proposed privacy-preserving sound and conventional sound masking.

	Privacy-preserving sound	Sound masking
Private information preserved by the sound	Speaker identity information included in recorded speech	Linguistic information contained in conversations
Target of the sound	Automatic speaker verification systems using recorded speech	People in the area where the user is

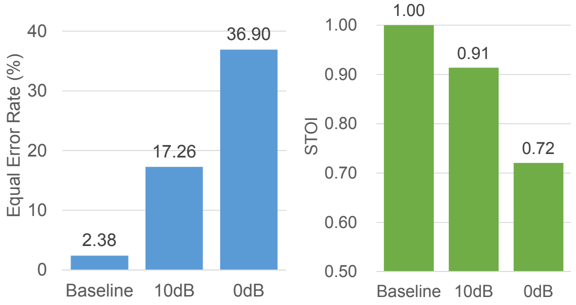

Fig. 2. Equal error rate (EER) and short-time objective intelligibility (STOI) with addition of white noise

Table 2. Band-pass filters used in the experiments

Bandwidth	Frequency ranges
1 kHz	0–1, 1–2, 2–3, 3–4, 4–5, 5–6, 6–7, 7–8 kHz
2 kHz	0–2, 1–3, 2–4, 3–5, 4–6, 5–7, 6–8 kHz
3 kHz	0–3, 1–4, 2–5, 3–6, 4–7, 5–8 kHz
4 kHz	0–4, 1–5, 2–6, 3–7, 4–8 kHz
5 kHz	0–5, 1–6, 2–7, 3–8 kHz
6 kHz	0–6, 1–7, 2–8 kHz
7 kHz	0–7, 1–8 kHz

4.1. Experimental conditions

The TIMIT speech database [15] was used in these experiments. Speech signals were sampled at 16 kHz and windowed at a 5-ms frame rate with a 25-ms window. The input feature was a 60-dimensional mel-frequency cepstrum coefficient (MFCC) vector, consisting of 19 MFCCs, 19 delta MFCCs, 19 delta-delta MFCCs, Energy, delta Energy, and delta-delta Energy. High-energy frames were retained and normalized so that the distribution of each cepstral coefficient had a mean of 0 and variance of 1 for a given utterance. A speaker verification system based on a Gaussian mixture model / universal background model (GMM-UBM) method was constructed by using the ALIZE 3.0 toolkit [16]. A 256-distribution UBM with a diagonal co-variance matrix was trained on 4620 utterances uttered by 326 male and 136 female speakers. Speaker-dependent GMMs were then estimated from the UBM by the maximum a posteriori (MAP) criterion. Speech data uttered by 112 male and 56 female speakers was used for estimating the speaker-dependent GMMs. The number of test utterances was 336.

The equal error rate (EER) was used as an objective measure to evaluate speaker verification performance, and the short-time objective intelligibility (STOI) was used as an objective measure to evaluate human speech intelligibility [17, 18]. The STOI outputs a score from 0 to 1, which correlates with intelligibility, i.e., a large score represents high intelligibility and a small score represents low intelligibility, by comparing natural speech with test speech.

4.2. Impact of the SNR on the EER and STOI

To analyze the impact of the SNR on the EER and STOI, white noise was added to the test speech so that the SNR for each utterance became 10 and 0 dB. The EER and STOI were then evaluated with the test speech, as shown in Figure 2. The results show that the EER increased with decreasing SNR. Therefore, a speaker's identity can be protected by creating a low SNR environment. The STOI, however, decreased with the SNR, meaning that speech intelligibility is largely degraded when noise is added to generate a low SNR. From these results, it is clear that the SNR reduction from white noise strongly impacts the EER and STOI, and that such noise would not work well as the desired privacy-preserving sound.

4.3. Impact of frequency on the EER and STOI

To analyze the impact of frequency on the EER and STOI, noise was applied using band-pass filters, which passe frequencies within a certain range and reject frequencies outside that range. Table 2 lists the conditions of the band-pass filters used in these experiments. Standard white noise and 35 different varieties of band-pass filtered white noise were added to the test data. In these experiments, the noise was added so that the SNR became 10 dB.

Figures 3 and 4 shows the results for the EER and STOI, respectively, for all the different sets of test data. Under 1-kHz bandwidth condition, Fig. 3 shows that the noise applied with 5–6 kHz band-pass filter gave the highest EER. Comparing the noise applied with filters passing 5 kHz, the noise applied with the 1–6 kHz band-pass filter gave the highest EER. Under each bandwidth condition, the EER was low with low-frequency noise, but high when the noise included the 5–6 kHz range. These results indicate that the frequency of noise strongly affects the EER, with the 5–6 kHz range having the strongest impact. Additionally, there was a trend that noise applied with a wide band-pass filter gave a high EER, and the 1–6 kHz filter gave the highest EER among all conditions, including standard white noise. Consequently, if an appropriate band-pass filter is chosen and sounds is created with the filter, the speaker verification performance can be efficiently degraded.

From Fig. 4, it can be seen that noise applied with the lower-frequency band-pass filters gave smaller STOI under all bandwidth conditions. This result corresponds to the fact that important phonetic information in speech is contained in low frequency. On the other hand, the difference in STOI across the range of band-pass filters was small. That is, the frequency of noise has a stronger affect on speech intelligibility than does the range of frequencies.

Comparing the results for the EER and STOI shown in Figs. 3 and 4, respectively, we can see that the sounds having the strongest

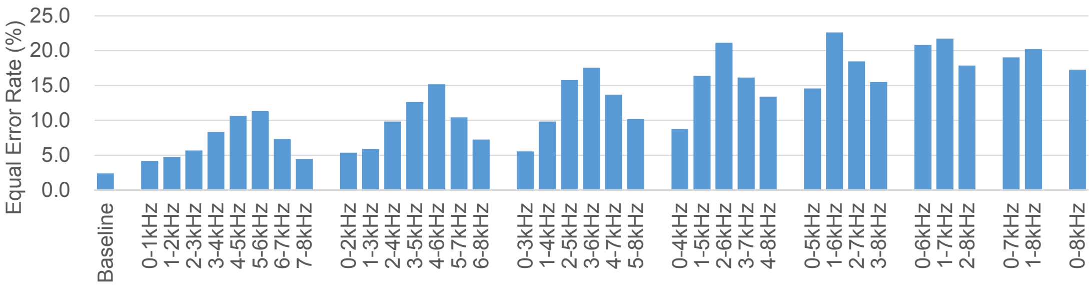

Fig. 3. EER with band-pass filtered noise (SNR: 10 dB)

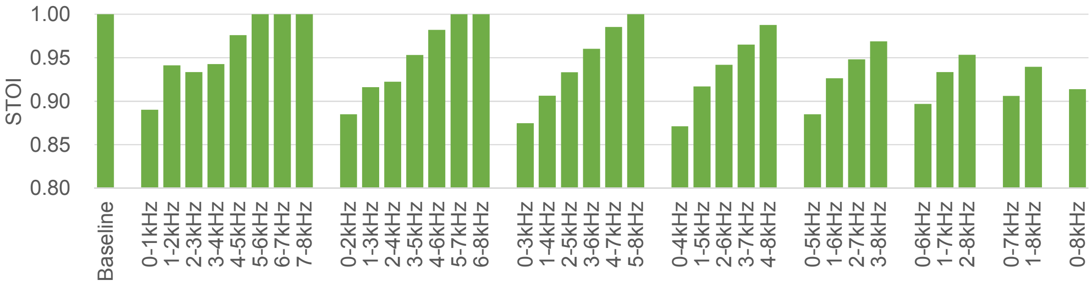

Fig. 4. STOI with band-pass filtered noise (SNR: 10 dB)

Table 3. Three best EER results achieving STOI larger than 0.914 (SNR: 10 dB). The EER and STOI for standard whit noise were 17.26% and 0.914.

	1–6 kHz	1–7 kHz	2–6 kHz
EER	22.62	21.73	21.13
STOI	0.926	0.934	0.942

impacts on the EER and STOI were different. By taking account of the difference, sound that degrades speaker verification performance without degrading human speech intelligibility can be created. Table 3 lists the three best EER results with STOI larger than 0.914. These cases all achieved higher intelligibility than did the case with standard white noise. At the same time, the results for the EER were degraded from the result for standard white noise. Therefore, noise with these characteristics can efficiently degrade speaker verification performance while maintaining human speech intelligibility.

4.4. Comparison of speakers

To compare the effect of noises for each speaker, the EER for each speaker was evaluated. In this experiment, the 1–6 kHz band-pass filtered noise was added to test data and the threshold for verification was tuned for each speaker. The resulted EERs were widely distributed, 0.0–61.1%. This result indicates that the impact on the speaker verification performance is strongly depend on the speaker. Therefore, in future, sound considering voice characteristics of enrolment speakers is required to protect identities of all speakers.

5. CONCLUSIONS

This paper has proposed privacy-preserving sound that can degrade the performance of automatic speaker verification systems without interfering with speech communication. The goal of our study is to prevent speaker identification from recorded speech by generating privacy-preserving sound in physical space. To make a first step toward solving this problem, we investigated what kind of sound would work well as privacy-preserving sound, through speaker verification experiments and objective evaluation of speech intelligibility. The experimental results show that appropriate sound can degrade speaker verification performance without degrading speech intelligibility. However, the degradation of speaker verification performance is not enough to protect private information and the sound still affect speech intelligibility. Consequently, the further improvement is necessary.

Future work will include experiments with speech database collected in a noisy environment, like that of the Speaker Recognition Evaluation (SRE) series conducted by the National Institute of Standards and Technology (NIST). Subjective evaluation will be also conducted in a real environment. In addition, we will investigate privacy-preserving sound considering voice characteristics of enrolment speakers.

6. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 15H01686 and the Telecommunications Advancement Foundation.

7. REFERENCES

[1] L. Cutillo and R. Molva, "Safebook: A privacy-preserving online social network leveraging on real-life trust," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 94–101, 2009.

[2] B. Debatin, J. Lovejoy, and A. Horn, "Facebook and online privacy: attitudes, behaviors, and unintended consequences," *Journal of Computer-Mediated Communication*, vol. 15, no. 1, pp. 83–108, 2009.

[3] 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.

[4] S. Furui, "Fifty years of progress in speech and speaker recognition," *Proceedings of 148th ASA Meeting 2004*, 2004.

[5] J.P. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[6] V. Hautamäki, T. Kinnunen, M. Nosratighods, K.A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," *Proceedings of Interspeech 2010*, pp. 1473–1476, 2010.

[7] R.G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.M. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," *Proceedings of Odyssey 2014*, pp. 137–144, 2014.

[8] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification by voice transformation," *Proceedings of ICASSP 2009*, pp. 3909–3912, 2009.

[9] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," *Proceedings of ASRU 2009*, pp. 529–533, 2009.

[10] S. H. K. Parthasarathi, M. Magimai-Doss, H. Bourlard, and D. Gatica-Perez, "Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios," *Proceedings of ICASSP 2010*, pp. 4474–4477, 2010.

[11] S. H. K. Parthasarathi, H. Bourlard, and D. Gatica-Perez, "Wordless sounds: robust speaker diarization using privacy-preserving audio representations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 85–98, 2013.

[12] K. Yamamoto, T. Masatoshi, and S. Nakagawa, "Privacy protection for speech signals," *Procedia Social and Behavioral Science*, vol. 2, no. 1, pp. 153–160, 2010.

[13] M. Akagi and Y. Irie, "Privacy protection for speech based on concepts of auditory scene analysis," *Proceedings of Internoise 2012*, p. 485, 2012.

[14] K. Ueno, H. Lee, S. Sakamoto, A Ito, M. Fujiwara, and Y Shimizu, "Experimental study on applicability of sound masking system in medical examination room," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3096, 2008.

[15] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[16] A. Larcher, J-F. Bonastre, B. Fauve, and K.A. Lee, "ALIZE 3.0 – Open source toolkit for state-of-the-art speaker recognition," *Processings of Interspeech 2013*, pp. 2768–2772, 2013.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *Proceedings of ICASSP 2010*, pp. 4214–4217, 2010.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.