INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An investigation of spoofing speech detection under additive noise and reverberant conditions

Xiaohai Tian[1,2], Zhizheng Wu [3], Xiong Xiao [4], Eng Siong Chng[1,2,4] and Haizhou Li[1,5]

[1]School of Computer Engineering, Nanyang Technological University (NTU),Singapore
[2]Joint NTU-UBC Research Center of Excellence in Active Living for the Elderly, NTU, Singapore
[3]The Center for Speech Technology Research, University of Edinburgh, United Kingdom
[4]Temasek Laboratories, NTU, Singapore
[5]Human Language Technology Department, Institute for Infocomm Research, Singapore
{xhtian, xiaoxiong, aseschng}@ntu.edu.sg, zhizheng.wu@ed.ac.uk, hli@i2r.a-star.edu.sg

Abstract

Spoofing detection for automatic speaker verification (ASV), which is to discriminate between live and artificial speech, has received increasing attentions recently. However, the previous studies have been done on the clean data without significant noise. It is still not clear whether the spoofing detectors trained on clean speech can generalise well under noisy conditions. In this work, we perform an investigation of spoofing detection under additive noise and reverberant conditions. In particular, we consider five difference additive noises at three different signal-to-noise ratios (SNR), and a reverberation noise with different reverberation time (RT). Our experimental results reveal that additive noises degrade the spoofing detectors trained on clean speech significantly. However, the reverberation does not hurt the performance too much.

Index Terms: Spoofing detection, noisy database, additive noise, reverberation, phase-based feature

1. Introduction

Recently, automatic speaker verification (ASV) has been significantly advanced to the point of mass-market adoption [1, 2, 3, 4]. However, most of current ASV systems assume human voices, and there are concerns that whether the systems can still achieve robust performance in the face of diverse spoofing attacks. A spoofing attack is that an attacker attempts to manipulate an ASV result for a target genuine speaker to obtain access permission. A significant amount of evidences have confirmed the vulnerability of current state-of-the-art ASV systems under spoofing attacks as reviewed in [5]. This has led to the active development of spoofing countermeasures, also called spoofing detection, that is to discriminate human and spoofed speech.

In the past several years, spoofing detection for speaker recognition has been studied on a variety of diverse datasets. In [6, 7], the Wall Street Journal (WSJ) corpus was used to assess countermeasures for speech synthesis attacks. In [8], the publicly available RSR2015 corpus was used to evaluate spoofing detection for replay attacks. In [9, 10], synthetic speech from the Blizzard challenge [11] was used for speech synthesis spoofing detection. In [12], a recently released spoofing and anti-spoofing (SAS) corpus as a standard spoofing database was used to assess speech synthesis and voice conversion spoofing countermeasures. We note that WSJ, SAS and Blizzard challenge databases were recorded by high-quality microphones in sound-proofing environment, while the RSR2015 corpus was recorded by multiple mobile devices in a quiet office room. All these databases do not have any significant channel and/or additive noise. These databases allow us to focus on spoofing effects but do not simulate practical scenarios of ASV applications.

There are also some studies that use data with channel noise. The National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) 2006 database which has significant telephone channel noise was used to assess voice conversion spoofing countermeasures in [13, 14, 15, 16, 17]. In [18], a so-called AVspoof database includes replay, speech synthesis and voice conversion spoofing attacks to simulate realistic scenarios, which re-recorded synthetic or voice-converted speech using multiple mobile devices.

In general, the databases used in the past spoofing detection studies seldom consider the noisy conditions[1], even the standard spoofing detection databases: SAS and ASVspoof 2015 challenge[2] databases. However, it is hard to avoid noise in real scenario. Hence, another concern for ASV deployment arises that whether currently developed spoofing detection algorithms/systems are still effective under noisy conditions.

In this work, our focus is the spoofing detection under additive noise and reverberant conditions. We perform a preliminary investigation of spoofing detection under noisy conditions using the state-of-the-art countermeasure techniques and then answer the following questions:

- Do current state-of-the-art spoofing detection algorithms work well under additive noise and reverberant conditions?
- How additive noises and reverberation noise affect the spoofing detection performance?
- What kind of noise is more serious to degrade the performance of spoofing detection algorithms?

We believe better understanding of above questions, and the noisy database will drive the development of generalised and noise robust spoofing detection algorithms.

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative and the DSO funded project MAISON DSOCL14045, Singapore.

[1]Preliminary studies under additive noise conditions have been done in [19, 20]. However, the reverberation noise is not considered in these papers. Moreover, [19] use the different classifier and features and also lead to different conclusions of this work.

[2]SAS corpus is available at: http://dx.doi.org/10.7488/ds/252 and ASVspoof 2015 corpus is available at: http://dx.doi.org/10.7488/ds/298

http://dx.doi.org/10.21437/Interspeech.2016-743

2. Noisy Database

In order to represent the practical application scenarios for spoofing detection and compensate to our previous noisy database [20], we generate a database in both additive noise and reverberant environments based on the ASVspoof 2015 challenge database [21]. The details and protocols about the ASVspoof database can be found in [21].

This noisy version aims to quantify the effects of current spoofing detection algorithms in additive noise and reverberant conditions. This database will also facilitate future assessment work in this task. In this section, we will briefly introduce the types of noise to be added, and the procedure of adding noise.

2.1. Noise signals

Five types of additive noises and three reverberation noise, representing the probable application scenarios, are used for the construction of the noisy ASVspoof database. Additive noises are chosen from two databases. White noise, speech babble and vehicle interior noise are selected from NOISEX-92 database [22]. While, street noise and cafe noise are selected from QUT-NOISE database [23]. These are standard types of additive noise widely used for speech recognition [24, 25, 23], speaker verification [26, 27] and speech enhancement [28]. We can classify these additive noises into stationary noise, including white noise and Volvo noise, and non-stationary noise, consisting of babble noise, street noise and cafe noise. The room impulse responses (RIRs) of the reverberation noises are simulated of a small size room. We briefly describe these noises as follows:

- **White Noise**: The random signal with a constant power spectral density.
- **Babble Noise**: Speech babble and the recording is made in a canteen with 100 people speaking.
- **Volvo Noise**: Vehicle interior noise and the recording is made in Volvo 340 on an asphalt road, rainy conditions.
- **Street Noise**: Mixed noise, which is made at the roadside near inner-city, mainly consisting of road traffic noise, pedestrian traffic noise and bird noise.
- **Cafe Noise**: Mixed noise, which is made in outdoor cafe environment, mainly consisting of speech babble and kitchen noise from the cafe environment.
- **Reverberation Noise**: It is a room impulse response (RIR) simulated with a small size room.

2.2. Adding noise

The data from ASVspoof database are considered as clean data. Noises are artificially added to the clean data. For each clean signal in the development and evaluation sets of ASVspoof database, eighteen noisy versions of the signal are generated, consisting of five additive noises in three SNR levels and reverberation noise with three reverberant times (RTs). The sampling frequency of all the data is 16 kHz.

To add the additive noise, the Filtering and Noise Adding Tool (FaNT)[3] is used. The noisy signals are generated by adding the clean speech and noise files together at various SNRs. As the silence periods appear in many speech files of ASVspoof database, it is important to calculate the SNR only based on the sections of speech signal. Given a clean signal, we take a segment of the noise signal with equal length as the clean signal but random starting point from the whole noise file. Then the noise segment is scaled and added to the clean signal in 20 dB, 10 dB and 0 dB SNR levels.

To add the reverberation noise, we convolve the clean signals with simulated RIRs. Three different acoustic conditions are simulated, considering the RT in 0.3 second, 0.6 second and 0.9 second.

After the adding noise process, the clipping may occur. In order to maintain a stable spectrogram representation of the signal, the signal is scaled to avoid the clipping.

3. Benchmarking system

In order to demonstrate the utility of the ASVspoof noisy database, we conduct a series of experiments to examine the performance of our spoofing speech detection system on both additive noise and reverberant scenarios.

The detection system consists of three parts, 1) the feature extraction module; 2) the classification module; and 3) the score fusion module.

3.1. Feature extraction

Similar to our previous system described in [29, 30, 20], six types of feature are extracted. Given a noisy waveform, the short-time Fourier transform (STFT) is applied on the speech signal using analysis window of 25ms with 15ms overlap. For the n-th frame, the magnitude and phase spectrum, $|\mathrm{X}(n,\omega)|$ and $\theta(n,\omega)$, are obtained by $\mathrm{X}(n,\omega) = |\mathrm{X}(n,\omega)|e^{j\theta(n,\omega)}$. Then, two magnitude-based features, namely log magnitude spectrum (LMS) and residual log magnitude spectrum (RLMS) are derived from $|\mathrm{X}(n,\omega)|$. Four phase-based features, namely instantaneous frequency derivative [31] (IF), baseband phase difference [32] (BPD), group delay [33] (GD) and modified group delay [33] (MGD), are derived from $\theta(n,\omega)$. The FFT length is chosen to be 512 and the dimension of all the original features are 256. Please find the detail information in [29, 30].

3.2. Classifier

The multilayer perceptron (MLP) based spoofing speech detection system [30, 20] is used in this work. Each of the features mentioned above with its delta and acceleration coefficients is used as the input vector to train its own classifier. The MLP, which contains one hidden layer with 2,048 sigmoid nodes, is used to predict the posterior probability of the input vector being synthetic speech. The score is calculated by averaging the posterior probabilities of all the frames over the utterance. Noted that, all the MLP classifier are trained from clean data.

3.3. Evaluation metrics and fusion

The equal error rate (EER) is used to evaluate the system performance.

As described in Section 3.1, different features are designed to detect different types of artifacts. In order to benefit the advantages of each feature and improve the system stability, a score level fusion is applied.

To avoid the over-fitting problem, the scores of all systems are simply averaged to produce the final score. The Bosaris toolkit[4] is used to compute the EERs of each feature and the fused system.

[3] http://dnt.kr.hs-niederrhein.de/

[4] https://sites.google.com/site/bosaristoolkit/

4. Experiments

4.1. Experimental setups

The dataset used in the experiments consist of three subsets, including training set, development set and evaluation set. To simplify the experiments, the training set is clean speech data taken from ASVspoof database. As the training set consists of clean data only, it models the speech without noise distortion and represents all the speech information. The best performance of the clean classifier is obtained in the case of testing on clean data, which can be found in our previous work [30].

The development and evaluation sets are chosen from the noisy ASVspoof database, including all the eighteen noise scenarios as described in Section 2. As the results on the development set are similar to that of the S1 to S5 on the evaluation set, we only report the results of the evaluation set. Because the classifier used in these experiments is the same as that of our previous work [30], these results are comparable with the results in clean condition.

4.2. Evaluation results

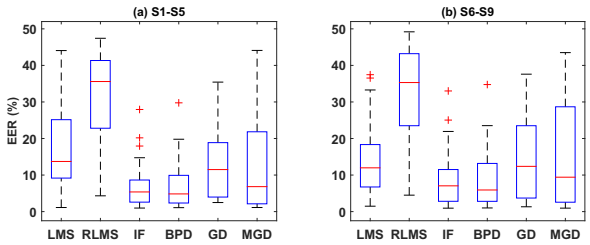

Figure 1: *Boxplot of averaged EERs (%) for different features on the noisy evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles.*

The results of different features on the noisy evaluation set are shown independently as the known attacks (S1-S5), the unknown attacks (S6-S9) and the unknown attacks generated by waveform concatenation (S10). For better analyse the performance of different feature across all the noisy conditions, boxplots are also provided in Figure 1. The boxplot for results of spoofing attack subsets, S1 to S5 and S6 to S9, are presented in Figure 1(a) and Figure 1(b), respectively. The detailed results are listed in Appendix A Table 1 .

We first compare the system performance across different types of attack under noise conditions. Figure 1(a) shows the system performance of S1-S5 attacks. Because such attacks are available for training, even in noisy condition, the lower error rates are obtained in these attacks than the rest types of attack. Although the error rates of S6-S9, as shown in Figure 1(b), is higher than that of S1-S5, the results still comparable. This is consistent with the results in clean condition [30]. For S10 attack, as shown in Table 1 of Appendix A, the error rates of all the features are significantly higher than that of S1-S9. Hence, we conclude that in both clean and noisy conditions, the detection of S10 is still the most challenge task among the spoofing attacks.

Then we analyse the effect of noisy data for the detection system using different features. In general, across the attacks S1-S9, the magnitude-based features, LMS and RLMS, perform worse than the phase-based features, IF, BPD, GD and MGD. In particular, in all the spoofing attacks, the RLMS obtains much higher EERs than other features. This may be due to that the LPC filter is not robust in noisy environments [34], which affects the quality of RLMS. Among the phase-based features, IF and BPD outperform other features in terms of the average EERs over all the noise scenarios. These observations are very different from [19], which report the magnitude-based features are more useful for spoofing detection under noisy scenarios.

As shown in Table 1 of Appendix A, some features are effective for particular noise scenario. For example, in the babble noise scenario, the MGD is capable to obtain low error rates. While in white, street, and cafe noises, IF and BPD perform much better than other features.

4.3. Fusion results

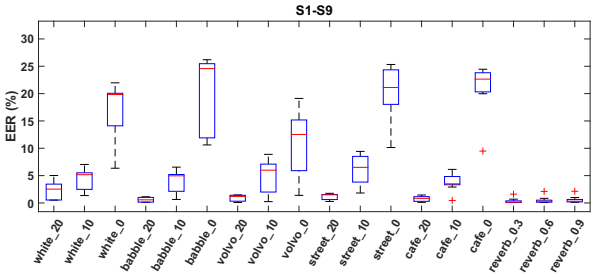

Figure 2: *Boxplot of fused EERs (%) for S1-S9 attacks under different noise scenarios of the evaluation set. Red lines are medians, box edges are at 25% and 75% quantiles.*

We first examine the results for S1-S9 attacks in additive noise and reverberation noise scenarios on the evaluation set. The boxplot of fused EERs for S1-S9 attacks under different noise scenarios are shown in Figure 2. From figure we observed that, for additive noise, the systems across all the noise scenarios perform worse than that of clean condition. The detection performance deteriorates as SNR decreases.

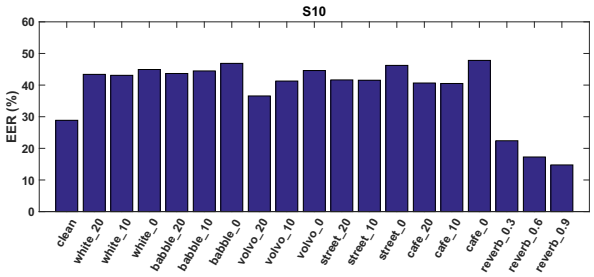

Figure 3: *Fused EERs (%) for S10 attacks under different noise scenarios of the evaluation set.*

Among all the additive noise scenarios, the system under Volvo noise scenario performs best, which constantly achieves lowest error rates. Especially, at SNR of 0 dB, the system under Volvo noise scenario outperform that of other noise scenarios significantly. This because the energy distribution of Volvo noise concentrates more on very low frequency (below 1 kHz), while the information in high frequency region is less distorted. This consistent with our previous results reported in [30], which confirms the effectiveness of the high-frequency region for synthetic speech detection. For the non-stationary noisy conditions, the features distorted by such noise are time-varying. Consequently, in these noise conditions, at SNR of 0 dB, the system performance degrades more than stationary noise scenar-

ios. This provides more challenge to the system to detect the spoofed attacks in such noisy conditions.

For reverberation noise scenarios, we found that the influence of reverberation noise for the system performance is very limited. Even use the RIR with different reverberation time, the system performs very stable. This may because the reverberant do not introduce new type of noise, the pattern still visible for the spoofing detector.

Then, we analyse the fused results of S10 attack in different noise scenarios on evaluation set. As shown in Figure 3, under additive noise scenarios, the system performs worse than that of clean condition. Alternatively, we surprisingly found the reverberation can significantly improve the performance on S10, and the performance further improve as the RT increase. It is may because that the reverberation performs temporal processing that helps to reveal the artefacts in features. For concatenated speech, there must be some discontinuity in the magnitude and phase. With the temporal filtering of reverberation, such discontinuity may become more obvious.

Finally, we analyse the averaged results across all the spoofing attacks. As shown in Table 2 of Appendix A, the systems performance heavily deteriorate under additive noise conditions. The detection performance decreases significantly as lower SNR. While, compare to clean condition, the systems performance are improved under reverberant scenarios.

5. Conclusions

In this paper, we constructed a noisy database, generated by adding both additive noise and reverberant scenarios, for spoofing and anti-spoofing research. To provide the benchmark results, the state-of-the-art spoofing detection system is used to detect the spoofing attacks in noisy conditions. The preliminary results using the classifier trained from clean data shown that,

- the systems performance degrade in all the additive noise scenarios and further deteriorate as SNR decreases;
- compare to the reverberation noise, the additive noises affect the detection system more seriously;
- the reverberation noise can improve the performance on concatenated speech;
- the system performance varies significantly under different noise scenarios and the phase-based features are more robust to noise than magnitude-based features.

In the future, we plan to exam the effectiveness of multi-condition training for spoofing detection under noisy conditions. We also going to investigate the reason why the reverberant is effective for S10 concatenated speech detection.

A. Appendix

Table 1: *Average EERs (%) of different features on the noisy evaluation set. Clean indicates the results of our previous work [30].*

	S1-S5 (Known)						S6-S9 (Unknown)						S10 (Unknown)					
Noise scenarios	LMS	RLMS	IF	BPD	GD	MGD	LMS	RLMS	IF	BPD	GD	MGD	LMS	RLMS	IF	BPD	GD	MGD
clean	**0.02**	**0.34**	**1.31**	**0.10**	**0.05**	**0.00**	**0.01**	**0.36**	**1.31**	**0.09**	**0.03**	**0.02**	**35.24**	**30.80**	**25.56**	**30.67**	**33.90**	**38.54**
white_20	11.57	22.81	4.11	3.07	8.58	7.55	7.53	23.50	5.40	3.76	7.77	9.17	47.80	49.05	41.30	39.54	48.35	47.82
white_10	24.65	30.35	7.15	5.65	14.95	27.54	17.39	29.32	10.29	7.82	14.61	28.68	49.54	49.69	42.05	39.86	48.75	48.79
white_0	39.26	36.87	17.94	14.43	30.20	44.10	37.46	33.88	25.04	21.48	31.76	43.50	49.80	49.85	44.17	42.75	49.43	49.28
babble_20	4.88	13.04	2.06	2.35	3.34	1.12	3.71	15.83	2.67	2.81	3.71	1.97	44.58	49.66	37.79	38.84	44.22	47.20
babble_10	15.23	25.78	8.64	9.92	15.07	5.09	12.63	25.93	11.51	13.18	17.26	9.00	45.16	49.86	42.44	42.95	46.33	48.39
babble_0	30.77	40.31	27.94	29.76	35.43	21.84	25.16	37.97	33.01	34.76	37.60	28.81	48.69	49.06	47.01	47.25	48.29	47.40
volvo_20	1.12	28.95	2.59	4.03	2.48	1.53	1.48	32.86	3.17	3.37	2.95	2.57	41.79	44.79	32.11	37.08	35.66	42.86
volvo_10	6.19	39.59	6.65	9.65	7.01	6.11	7.54	46.75	8.70	9.72	8.59	9.66	44.30	46.57	37.34	42.77	39.29	46.14
volvo_0	13.31	44.86	14.74	17.76	14.91	15.50	14.71	49.19	18.21	20.26	18.54	20.55	46.22	47.34	42.58	47.60	44.06	47.82
street_20	9.17	35.16	2.98	2.90	7.41	3.47	6.74	37.73	3.58	3.46	10.37	6.42	42.61	49.18	40.68	40.53	47.10	46.83
street_10	25.15	41.34	7.42	6.71	18.87	13.21	18.34	42.23	9.00	8.83	23.50	19.68	43.65	49.07	42.86	42.96	47.88	47.48
street_0	41.04	46.87	20.18	19.80	33.49	30.42	33.28	46.85	21.93	23.52	35.82	35.44	48.59	47.54	45.59	44.98	48.77	47.84
cafe_20	10.67	36.02	2.66	2.15	6.43	2.61	6.44	36.74	2.82	2.30	6.91	3.91	42.28	48.78	40.54	39.45	48.10	46.98
cafe_10	23.44	43.45	3.85	3.39	14.42	11.68	14.75	43.21	4.33	4.01	14.39	14.36	43.22	47.13	40.38	40.57	48.15	47.92
cafe_0	44.06	47.42	8.03	7.88	27.25	33.40	36.54	47.57	9.07	10.23	24.76	33.12	49.94	44.84	42.22	42.37	48.49	48.78
reverb_0.3	5.44	4.97	1.90	1.86	3.69	1.49	3.72	4.82	1.82	1.73	1.83	0.95	37.41	30.03	24.51	22.58	27.02	30.11
reverb_0.6	11.62	4.77	1.31	1.34	3.56	1.98	8.92	4.74	1.19	1.21	1.32	1.29	39.93	23.18	19.47	18.05	19.72	24.89
reverb_0.9	14.13	4.32	0.97	1.08	3.98	2.13	11.32	4.50	0.94	1.00	1.58	1.44	41.46	18.73	16.59	15.95	17.91	22.38

Table 2: *EERs (%) of fused system on both development and evaluation sets. Clean indicates the results of our previous work [30].*

	Development						Evaluation										
Noise scenarios	S1	S2	S3	S4	S5	Average	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Average
clean	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**28.87**	**2.89**
white_20	2.30	3.43	0.54	0.49	3.81	2.11	3.42	2.53	0.52	0.52	3.51	2.82	5.02	0.54	2.39	43.39	6.47
white_10	3.65	5.92	1.76	1.70	6.13	3.83	4.53	5.23	1.47	1.35	5.92	5.38	7.04	2.82	5.17	43.09	8.20
white_0	13.92	18.69	6.42	6.15	17.93	12.62	16.60	20.30	6.65	6.35	19.98	19.90	21.97	18.50	19.83	44.93	19.50
babble_20	0.61	1.27	0.13	0.13	1.36	0.70	1.05	0.94	0.13	0.15	0.15	1.00	1.13	0.26	0.57	43.67	5.05
babble_10	4.36	5.34	0.98	0.87	4.89	3.29	4.96	5.14	0.64	0.65	5.40	5.12	6.55	2.62	3.00	44.48	7.86
babble_0	24.88	25.56	11.47	11.25	24.70	19.57	24.95	25.55	10.61	10.61	25.45	26.20	24.57	12.34	20.26	46.87	22.74
volvo_20	0.83	0.93	0.01	0.00	0.98	0.55	1.21	1.19	0.09	0.09	1.40	1.47	0.90	1.40	0.40	36.57	4.47
volvo_10	7.79	6.05	0.17	0.18	4.36	3.71	8.89	7.32	0.24	0.29	5.45	6.09	6.00	7.03	2.56	41.28	8.52
volvo_0	19.03	14.09	1.50	1.50	10.26	9.27	19.12	15.48	1.39	1.39	11.30	12.53	15.08	14.15	7.41	44.60	14.24
street_20	1.35	1.76	0.35	0.33	1.38	1.03	1.51	1.76	0.27	0.28	1.49	1.61	1.53	1.00	0.73	41.64	5.18
street_10	6.95	8.87	3.98	3.85	7.93	6.32	7.33	9.04	3.84	3.74	8.34	9.43	6.51	1.82	6.47	41.54	9.81
street_0	21.55	23.17	17.43	17.32	23.03	20.50	23.35	24.30	18.11	17.80	24.50	25.33	21.08	10.14	21.12	46.23	23.20
cafe_20	1.15	1.13	0.53	0.49	1.42	0.95	1.05	1.06	0.30	0.29	1.45	1.43	0.84	0.11	0.65	40.66	4.78
cafe_10	4.06	5.15	4.35	4.24	5.83	4.73	3.56	4.59	3.47	3.50	5.56	6.14	2.91	0.46	3.50	40.51	7.42
cafe_0	19.63	20.60	22.04	21.92	21.78	21.20	22.05	22.65	24.01	23.75	23.70	24.46	19.98	9.49	20.43	47.81	23.83
reverb_0.3	0.30	0.47	0.12	0.18	1.99	0.61	0.25	0.28	0.14	0.12	1.61	0.69	0.10	0.08	0.08	22.41	2.58
reverb_0.6	0.18	0.35	0.22	0.23	2.32	0.66	0.27	0.37	0.40	0.32	2.11	0.85	0.10	0.15	0.09	17.27	2.19
reverb_0.9	0.17	0.46	0.32	0.34	2.37	0.73	0.26	0.52	0.40	0.32	2.14	0.99	0.12	0.21	0.15	14.76	1.99

B. References

[1] K. A. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," in *IEEE Signal Processing Society Speech and language Technical Committee Newsletter*, February 2013.

[2] M. Khitrov, "Talking passwords: voice biometrics for data access and security," *Biometric Technology Today*, vol. 2013, no. 2, pp. 9–11, 2013.

[3] B. Beranek, "Voice biometrics: success stories, success factors and what's next," *Biometric Technology Today*, vol. 2013, no. 7, pp. 9–11, 2013.

[4] W. Meng, D. Wong, S. Furnell, and J. Zhou, "Surveying the development of biometric user authentication on mobile phones," *IEEE Communications Surveys and Tutorials*, 2015.

[5] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[6] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[7] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[8] Z. Wu, S. Gao, E. S. Chng, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2014.

[9] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *INTERSPEECH*, 2012.

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

[11] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, p. e006, 2014.

[12] Z. Wu, P. L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[13] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Interspeech 2012*, 2012.

[14] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proc. Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012.

[15] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[16] E. Khoury, T. Kinnunen, A. Sizov, Z. Wu, and S. Marcel, "Introducing i-vectors for joint anti-spoofing and speaker verification," in *Proc. Interspeech*, 2014.

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

[18] S. K. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015.

[19] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," in *http://arxiv.org/pdf/1603.03947v1.pdf*, 2016.

[20] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "spoofing detection under noisy conditions: A preliminary investigation and an initial database," in *http://arxiv.org/pdf/1602.02950v1.pdf*, 2016.

[21] 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*, 2015.

[22] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," in *Tech. Rep., DRA Speech Research Unit*, 1992.

[23] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Proc. INTERSPEECH*, 2015.

[24] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[25] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[26] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1998.

[27] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[28] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[29] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for ASVspoof 2015 challenge," in *Proc. INTERSPEECH*, 2015.

[30] X. Tian, Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Spoofing detection from a feature representation perspective," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[31] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[32] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[33] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2281–2289, 1992.

[34] S. M. Kay, "The effects of noise on the autoregressive spectral estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 5, pp. 478–485, 1979.