Human vs Machine Spoofing Detection on Wideband and Narrowband Data

Mirjam Wester[1], *Zhizheng Wu*[1], *Junichi Yamagishi*[1,2]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2] National Institute of Informatics, Japan

m.wester@inf.ed.ac.uk, zhizheng.wu@ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

How well do humans detect spoofing attacks directed at automatic speaker verification systems? This paper investigates the performance of humans at detecting spoofing attacks from speech synthesis and voice conversion systems. Two speaker verification tasks, in which the speakers were either humans or machines, were also conducted. The three tasks were carried out with two types of data: wideband (16kHz) and narrowband (8kHz) telephone line simulated data. Spoofing detection by humans was compared to automatic spoofing detection (ASD) algorithms. Listening tests were carefully constructed to ensure the human and automatic tasks were as similar as possible taking into consideration listener's constraints (e.g., fatigue and memory limitations). Results for human trials show the error rates on narrowband data double compared to on wideband data. The second verification task, which included only artificial speech, showed equal overall acceptance rates for both 8kHz and 16kHz. In the spoofing detection task, there was a drop in performance on most of the artificial trials as well as on human trials. At 8kHz, 20% of human trials were incorrectly classified as artificial, compared to 12% at 16kHz. The ASD algorithms also showed a drop in performance on 8kHz data, but outperformed human listeners across the board.

Index Terms: spoofing, human performance, automatic spoofing detection

1. Introduction

Due to the development of channel and noise compensation techniques the accuracy of automatic speaker verification (ASV) systems has advanced significantly in recent years to the point of mass-market adoption [1]. However, a major challenge in the deployment of ASV systems is dealing with spoofing attacks. A spoofing attack is when an attacker attempts to manipulate a verification result by mimicking a client speaker in person or by using some advanced technologies, such voice conversion or speech synthesis.

As identified in [2], there are at least four types of spoofing attacks: impersonation [3, 4], replay [5], speech synthesis [6, 7] and voice conversion [8, 9, 10, 11]. Recently, due to the development of speech synthesis and voice conversion technologies, a number of off-the-shelf open-source toolkits have become available. Hence, speech synthesis (SS) and voice conversion (VC) have become two of the most easily accessible and effective techniques to carry out spoofing attacks [12, 2], and constitute a serious risk to ASV systems.

The main aim of the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge at Interspeech 2015 is to test how vulnerable (or robust) ASV systems are to speech synthesis or voice conversion spoofing attacks. This paper addresses human performance on this task. The question addressed here is how well humans perform at identifying human impostors and artificial impostors and how well they are able to detect spoofing attacks. The database, **SAS** [13], used in the ASVspoof Challenge includes 16kHz data for text-independent speaker verification. The 16kHz sampling rate was for the benefit of the synthetic and voice conversion techniques used in the spoofing attacks. However, a more realistic scenario in the ASV world is of speech that is transmitted through a telephone line, i.e., narrowband data. Therefore, in this study we include 8kHz data and we investigate how the human performance changes when the data is narrowband telephone style speech instead of wideband speech at 16kHz sample rate. Human performance is compared to the performance of automatic spoofing detection (ASD) algorithms on the spoofing detection task at both 8 and 16 kHz.

Various studies over the years have shown that the performance of machines is becoming equal to or even surpassing humans on certain speaker verification tasks. In [14] the speaker verification performance of human listeners was compared to that of ASV systems on the NIST 1998 Speaker Evaluation Data. The results showed ASV systems performed really well, but under degraded conditions human performance was more robust. Similarly, Wenndt and Michell [15] compared human recognition vs machine recognition for changing environments, e.g., short sentences, frequency selective noise and time-reversed speech and for most conditions they found that humans were more robust. More recently studies have shown that ASV systems are performing as well as listeners even under degraded conditions. Hautamäki and colleagues [16] showed on NIST SRE 2008 data that their joint factor analysis (JFA) [17] ASV system outperformed human listeners on an *easy* dataset in which there was minimal channel mismatch. On the *hard* dataset, which included severely mismatched channel conditions, the JFA system was found to perform as well as humans. However, it should be noted that their listeners were all non-native, who underperform on many listening tasks compared to native listeners [18, 19, 20].

How well humans perform at detecting spoofing attacks has not been studied extensively, only a handful of papers [21, 22] have addressed the detection of human imitators by both machines and humans. Hautamäki et al. [21] found that automatic systems make less errors than humans when evaluating a person who is intentionally modifying their voice and Zetterholm et al. [22] concluded from their study that ASV systems and humans evaluate imitations differently. To our knowledge, no studies address human performance on SS and VC spoofing attacks. The lack of insight regarding the performance of humans on the task of spoofing detection motivates the current paper.

The next section sets out how the human evaluation was carried out and describes the automatic spoofing detection. This

is followed by results of the human and automatic performance as well as a short discussion of these results.

2. Method

The experimental design we adopted in this study and how elements of the design were motivated by a pilot test are described below. This is followed by details of the three human listening tasks. Next, the automatic spoofing detection system, the spoofing data materials and the spoofing systems are described.

2.1. Pilot experiment

Our listening experiments were set up keeping the constraints of machine and human in mind. As described in [14] many of the rules, for example, of the NIST evaluation could not be applied equally to people and machines. Things like listener fatigue, boredom and memory limitations play a role for humans but not machines, whereas humans have the advantage of having heard speech from the day they were born or even before then, in utero. Consequently –in our pilot test– we assumed that because humans hear huge amounts of speech on a daily basis their performance, when detecting synthetic or voice converted speech, would quickly reach ceiling levels. However, our pilot test (20 listeners) revealed that this detection task was more difficult for human listeners than we had thought, in first instance.

The pilot experiment showed no obvious ceiling effect. The experiment consisted of 260 stimuli: 130 artificial samples (13 systems with 10 samples each) and 130 human samples, all at 16kHz and randomly selected. Listeners were instructed to judge –for each sample– whether the sample was from a human or a machine. The overall error rate for the systems came to an average of 31.6% of artificial samples classified as human (min 15.5% and max 72%) and 10.6% of natural samples classified as artificial. These error rates were much higher than we expected. We hypothesised this was due to a mismatch between a listener's mental representation of speech and the type of speech they were hearing in the experiment. For instance, the samples are very short (2-3 sec) and the recording conditions are such that some samples may result in being classified as a distorted human voice rather than synthetic speech. On the basis of this pilot, we refined the listening test to include training material, i.e., by letting a subject hear examples of the recordings we expect their mental representation to become more attuned to the task, thus enabling the subject to judge the samples more accurately. Additionally, we extended the instructions and included a role playing element to encourage listeners to perform the task to the best of their ability.

2.2. Human listening tests

We conducted three human listening tests: two verification tasks and one detection task for two different types of conditions: 16kHz data and narrowband telephone line simulated data, 8kHz. The first verification task contained only human samples, the second verification task contained human training samples but all test samples were artificial (SS or VC). The third task –the detection task– contained both human and artificial samples and the goal for the listener was to correctly detect whether the sample was produced by a human or a machine.

2.2.1. Listeners

Experiments were carried out using a web interface. In total, 100 native English listeners took part in the 16kHz experiment and 30 in the 8kHz experiment. The results presented in this paper include only the first 30 listeners of the 16kHz data experiment as they are directly comparable to the 30 listeners in the 8kHz experiment. Listeners were seated in a sound isolated booth and listened to all samples using Beyerdynamic DT 770 PRO headphones. Each listener did all three tasks. On average it took about an hour to complete the experiment. Listeners were remunerated for their time and effort.

2.2.2. Task 1: Speaker verification (human)

In the human speaker verification task the listeners were asked to imagine they were responsible for giving people access to their bank accounts. They were informed that they would only have a short recording of a person's voice to base their judgement on. It was stressed that it was important to not give access to "impostors" but equally important that access was given to the "bank account holder".

The listeners were given five sentences from each target speaker to familiarise themselves with the voice. After listening to the training samples they were given 21 trials to judge as SAME or DIFFERENT. The trials were pairs of samples; a reference sample and the test sample. This was repeated for three different target speakers.

In total, 46 target speakers (20 Male, 26 Female) were rated. Each target speaker was judged by two listeners. The number of target vs non-target varied per speaker to keep listeners from keeping count for individual speakers. On average there were 10 targets and 11 non-targets per speaker. Genders were not mixed within a trial.

2.2.3. Task 2: Speaker verification (artificial)

In the second task, listeners were asked to decide whether an artificial voice[1] sounded like the original speaker's voice. The listeners were informed that the artificial voice would sometimes sound quite degraded but were asked to ignore the degradations as much as possible. Additionally, they were told that there would be artificial voices that were supposed to sound like the intended speaker as well as artificial voices that were not supposed to match the original speaker. The challenge was framed as "your challenge is to decide which of the artificial voices are based on the "bank account holder's voice" and which are based on an "impostor's voice"."

As in the first task, the listeners were given five natural speech samples from the intended speaker to familiarise themselves with the voice. After listening to the training samples, subjects were presented with pairs of reference and test samples to judge as SAME or DIFFERENT. It was made clear to the listeners that the test sample would be of an artificial voice. The reference sample was always natural speech.

This second task covered 46 target speakers in total. Each target speaker was judged by two listeners. For each target speaker there were 65 trials (13 systems, each presented 5 times). On average there were 39 targets and 26 non-targets per speaker. Once again gender was not mixed within any of the trials.

2.2.4. Task 3: Detection

In the final task, listeners were asked to judge whether a speech sample was a recording of a human voice, or a sample of an artificial voice. The challenge to the listeners was formulated

[1] Artificial was explained to the listeners as being "produced by a machine, computer-generated, for example a synthetic voice".

as: "Imagine an impostor trying to gain access to a bank account by mimicking a person's voice using speech technology. You must not let this happen. Your challenge in this final section is to correctly tell whether or not the sample is of a HUMAN or of a MACHINE."

For this final task, the listeners were also given some training samples. They listened to five samples of human speech recordings from one speaker (not present in the detection task) and five examples of artificial speech generated using five different methods (again the "speakers" were not in the test but the methods were). Finally, the listeners were informed that the training samples did not cover all the types of artificial speech.

In Task 3, there were 130 samples (65 human, 65 artificial (13 x 5)), and those samples were randomly selected from the evaluation set for each listener.

2.3. Automatic Spoofing Detection (ASD) system

One fused ASD system was used to compare automatic and human spoofing detection results. The fused system is a combination of Mel-frequency cepstral coefficients (MFCCs) and cosine-normalised phase (cos-phase) feature based detectors. Both MFCCs and cos-phase features include 18 dimensional static features, their deltas and delta deltas. The reason for choosing these two feature sets is that they are easy to extract without tuning hyper-parameters like, e.g., modified group delay features [10]. More details on the cos-phase features can be found in [23]. As for the classifier, we used a simple Gaussian mixture model with universal background model (GMM-UBM) based classifier with 1024 Gaussian components. Fusion was implemented using the BOSARIS Toolkit[2] at the score level.

2.4. Materials

The materials for the listening test were selected from Part-E of the spoofing database **SAS** [13, 24]. **SAS** contains speech data from 45 male and 61 female speakers selected from the Voice Cloning Toolkit (VCTK) database. The data in **SAS** is divided into five parts:

- **Part-A:** 24 parallel utterances (i.e., same across all speakers) per speaker: training data for spoofing algorithms.
- **Part-B:** 20 non-parallel utterances per speaker: additional training for spoofing algorithms.
- **Part-C:** 50 non-parallel utterances per speaker: enrolment data for client model training in speaker verification, or training data for speaker-independent countermeasures.
- **Part-D:** 100 non-parallel per speaker: development set for speaker verification and countermeasures.
- **Part-E:** Around 200 non-parallel utterances per speaker: evaluation set for speaker verification and countermeasures.

2.4.1. Telephone channel

The 16kHz data was downsampled to 8kHz and then filtered with the G.712 frequency characteristic as defined by ITU for telephone equipment. We used the FaNT simulation tool [25] to filter the speech.

2.5. Spoofing systems

Five speech synthesis (SS) and eight voice conversion (VC) systems were developed for spoofing attacks. Most of the systems have been described in more detail in [13]. Here it suffices to mention the most salient details of the systems.

SS-LARGE-16: HMM-based TTS system [26]. The average voice is trained on the voice bank corpus [27] which includes hundreds of English speakers. This average voice is adapted using the target speaker's 16kHz data from Part-A and Part-B.
SS-LARGE-48: Same as SS-LARGE-16, except for adapted using 48 kHz data.
SS-SMALL-16: Same as SS-LARGE-16, except for using only Part-A adaptation data.
SS-SMALL-48: Same as SS-SMALL-16, except for adapted using 48 kHz data.
SS-MARY: Unit-selection system implemented by the MARY text-to-speech system (MaryTTS)[3] based on 16 kHz data from Part-A and Part-B.
VC-C1: Voice conversion with modified spectral slope. First coefficient of the source speaker's Mel-Cepstral coefficients (MCCs) was shifted.
VC-EVC: A many-to-many eigenvoice conversion (EVC) system [28]. Training data was taken from Japanese databases (ATR and JNAS). Conversion function only applied to MCCs.
VC-FEST: GMM-based voice conversion using Festvox.
VC-FS: Frame selection voice conversion system, simplified version of exemplar-based unit selection [29]. Only MCCs were converted.
VC-GMM: An enhanced version of VC-FEST GMM-based voice conversion.
VC-KPLS: Voice conversion using kernel partial least square (KPLS) regression [30].
VC-TVC: Tensor-based arbitrary voice conversion (TVC) [31]. The same Japanese dataset as in VC-EVC was used.
VC-LSP: GMM-based voice conversion with line spectral pairs and delta coefficients as the spectral features.

3. Results

3.1. Task 1: Speaker verification (human)

Table 1 presents the verification error rates for task 1 –the human speaker verification task– at 8 and 16kHz. On the 16kHz evaluation data, listeners identified impostors as genuine targets 6.26% of the time (FAR) while 1.18% of genuine trials were misclassified as impostors (FRR). For the telephone channel simulation at 8kHz the rate at which impostors are identified as genuine targets increases to 13.33% and the misclassification of impostors increases to 3.94%.

	8kHz	16kHz
Genuine (FAR):	13.33	6.26
Impostor (FRR):	3.94	1.18

Table 1: *Task 1 – Speaker verification (human) – human listeners' error rates in percentages.*

3.2. Task 2: Speaker verification (artificial)

Table 2 shows the acceptance rate of synthetic speaker verification. In this case, a genuine trial is a trial which was synthesised using the target speaker's voice while an impostor trial is a trial that was synthesised using a non-target speaker's voice. A higher acceptance rate indicates that the artificial system (SS or VC) is recognised more as the target speaker, i.e., it gives an indication of how well the artificial system imitates the target, or in other words, how similar the SS or VC system is to the target.

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

[3] http://mary.dfki.de/

Overall the SS systems achieve higher acceptance rates than the VC systems. SS-MARY (unit-selection system) results in the highest acceptance rate, while VC-C1 (modified spectral slope) achieves the lowest acceptance rate. There is not a great deal of difference between the results on 8kHz and 16kHz data with some systems leading to increases in acceptance rate, whereas others result in decreases, e.g., two out of five SS systems and five out of eight VC systems show a reduction in the acceptance rate when going from 16kHz to 8kHz data.

	8kHz	16kHz
SS-SMALL-16	32.22	31.94
SS-SMALL-48	34.81	28.70
SS-LARGE-16	34.81	38.89
SS-LARGE-48	35.19	32.87
SS-MARY	68.15	74.54
VC-GMM	27.78	32.87
VC-KPLS	23.33	31.48
VC-TVC	25.19	19.91
VC-EVC	21.85	23.15
VC-FS	40.37	38.43
VC-C1	10.00	6.94
VC-FEST	28.15	29.17
VC-LSP	21.48	25.93

Table 2: *Task 2 –Speaker verification (artificial) – human listeners' acceptance rate in percentages.*

3.3. Task 3: Detection

The spoofing detection results are presented in Table 3. Human listeners and the ASD algorithm judged samples and labelled them as either HUMAN or MACHINE. The percentages in Table 3 denote error rates which show the amount of times a human sample is misclassified as artificial or vice versa an artificial sample misclassified as human.

For human subjects, there is an increase in detection error rate when going from 16kHz to 8kHz data. On human trials, the error rate increases from around 12% to 20%, i.e. at 8kHz a fifth of the human trials are misclassified as artificial. On artificial trials, the overall error rate for spoofing detection goes from 8% to 12%. The system most successful at fooling human listeners is VC-C1 and the least successful spoof is VC-FS. This can be explained by the fact that VC-C1 only changes the slope of the spectral envelope and hence preserves naturalness whereas VC-FS generates target speech by selecting frames, and introduces significant artefacts due to the discontinuity across frames.

In general, the ASD system achieves much lower error rates than the humans except for on SS-MARY, a unit-selection system which directly uses waveforms to generate the spoofed speech. Interestingly, in the narrowband condition, ASD performs considerably better on SS-MARY than in the wideband condition.

4. Discussion

Human verification error rates double on the verification task which includes only human samples when reducing the sampling rate from 16 to 8kHz. This is unsurprising and probably in line with what one would expect. It can be compared to, for instance, the difficulty sometimes encountered when trying to tell apart siblings on the phone by their voice, or alternatively mother-daughter or father-son pairs.

	8kHz human	16kHz human	8kHz ASD	16kHz ASD
Human	19.64	12.48	1.03	0.30
All spoof	12.46	8.43	6.08	7.82
SS-SMALL-16	2.67	8.89	0.00	0.00
SS-SMALL-48	14.00	2.96	0.00	0.00
SS-LARGE-16	8.00	3.70	0.00	0.00
SS-LARGE-48	8.67	8.15	0.00	0.00
SS-MARY	6.00	8.15	48.14	98.08
VC-GMM	20.00	17.04	13.16	1.36
VC-KPLS	12.67	4.44	0.18	0.00
VC-TVC	14.67	5.19	0.04	0.03
VC-EVC	7.33	8.15	0.01	0.06
VC-FS	6.00	2.96	0.09	0.00
VC-C1	40.67	27.41	4.76	1.18
VC-FEST	10.67	5.19	12.59	0.73
VC-LSP	10.67	7.41	0.04	0.24

Table 3: *Task 3 –Spoofing detection– human and ASD detection error rate.*

On the second verification task, which consisted only of artificial samples, we don't see the same increase in verification error rate when going from 16kHz down to 8kHz. This task can be seen as a measure of the similarity of a SS or VC voice to the original target voice. Seen in this light, the SS systems outperform VC systems in almost all cases, i.e., they achieve better similarity to target speakers. In future work, we will further investigate the correlation between human subjects' scores and ASV scores on this task. If there is a correlation, ASV may be used as a tool to automatically measure speaker similarity for speech synthesis and voice conversion.

Compared to the pilot test results, the spoofing detection of artificial samples was much better in the current study. This indicates that the training was beneficial. However, the results on human samples is slightly worse here than in the pilot. Possibly, the listeners were erring too much on the side of caution as a result of the instructions. The ASD algorithm clearly outperforms human listeners on all systems except for the unit-selection voice SS-MARY. We can conclude that ASD systems and humans detect spoofing differently, in a similar vain to what [22] found for imitations.

With SS and VC techniques available off-the-shelf, creating artificial voices to spoof ASV systems is increasingly easy and constitutes a real threat. The ASVspoof challenge was set up to test how vulnerable (or robust) ASV systems are to SS or VC spoofing attacks. This paper investigated how well humans perform at identifying human and artificial impostors and how well they are able to detect spoofing attacks on 8kHz and 16kHz data. We found that error rates increase when rating human impostors at 8kHz rather than 16kHz.The acceptance rate of artificial samples did not vary much due to sampling rate. On the whole, SS systems scored higher, i.e., more similar to the targets than VC systems. Regarding spoofing detection rates, there is a drop in performance when going from 16kHz to 8kHz and the ASD algorithm outperforms humans on all systems, except for on SS-MARY.

All research data associated with this paper can be found at Edinburgh DataShare (`http://hdl.handle.net/10283/790`) [32].

Acknowledgements This work was partially supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology).

5. 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*, 2013.

[2] 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.

[3] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symposium on Intelligent Multimedia, Video and Speech Processing*, 2004.

[4] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry," in *Proc. Interspeech*, 2013.

[5] 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.

[6] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010.

[7] 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.

[8] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates," in *Proc. Interspeech*, 2007.

[9] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[10] 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.

[11] Z. Kons and H. Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems," in *Proc. Interspeech*, 2013.

[12] N. Evans, T. Kinnunen, J. Yamagishi, Z. Wu, F. Alegre, and P. DeLeon, "Voice anti-spoofing," in *Handbook of biometric anti-spoofing*, S. Marcel, S. Z. Li, and M. Nixon, Eds. Springer, 2014.

[13] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, and S. King, "SAS: A speaker verification spoofing database containing diverse attacks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.

[14] A. Schmidt-Nielsen and T. H. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data," *Digital Signal Processing*, vol. 10, no. 1, pp. 249–266, 2000.

[15] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 4245–4248.

[16] V. Hautamäki, T. Kinnunen, M. Nosratighods, K. A. Lee, B. Ma, and H. Li, "Approaching human listener accuracy with modern speaker verification," in *Proc. Interspeech*, 2010.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[18] A. R. Bradlow and D. B. Pisoni, "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2074–2085, 1999.

[19] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[20] A. Cutler, M. L. G. Lecumberri, and M. Cooke, "Consonant identification in noise by native and non-native listeners: Effects of local context," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1264–1268, 2008.

[21] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014, pp. 137–144.

[22] E. Zetterholm, M. Blomberg, and D. Elenius, "A comparison between human perception and a speaker verification system score of a voice imitation," in *Proc. of Tenth Australian Int. Conf. on Speech Science & Technology*, 2004.

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

[24] Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, Z.-H. Ling, and S. King, "Spoofing and Anti-Spoofing (SAS) corpus v1.0 [dataset]," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2015, http://dx.doi.org/10.7488/ds/252.

[25] H.-G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems." in *Proc. Interspeech*, 2005, pp. 2697–2700.

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

[27] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA*, 2013.

[28] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4822–4825.

[29] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *Proc. Interspeech*, 2013.

[30] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[31] D. Saito, N. Minematsu, and K. Hirose, "Effects of speaker adaptive training on tensor-based arbitrary speaker conversion," in *Proc. Interspeech*, 2012.

[32] M. Wester, Z. Wu, and J. Yamagishi, "Human vs machine spoofing [dataset]," University of Edinburgh, 2015, http://dx.doi.org/10.7488/ds/258.