ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge

Zhizheng Wu[1] Tomi Kinnunen[2] Nicholas Evans[3] Junichi Yamagishi[1]
Cemal Hanilçi[2] Md Sahidullah[2] Aleksandr Sizov[2]

[1]University of Edinburgh, United Kingdom
[2]University of Eastern Finland, Finland
[3]EURECOM, France

zhizheng.wu@ed.ac.uk, tomi.kinnunen@uef.fi, evans@eurecom.fr, jyamagis@inf.ed.ac.uk

Abstract

An increasing number of independent studies have confirmed the vulnerability of automatic speaker verification (ASV) technology to spoofing. However, in comparison to that involving other biometric modalities, spoofing and countermeasure research for ASV is still in its infancy. A current barrier to progress is the lack of standards which impedes the comparison of results generated by different researchers. The ASVspoof initiative aims to overcome this bottleneck through the provision of standard corpora, protocols and metrics to support a common evaluation. This paper introduces the first edition, summaries the results and discusses directions for future challenges and research.

Index Terms: Speaker verification, Spoofing, Anti-spoofing, Countermeasure, Spoofing detection

1. Introduction

Automatic speaker verification (ASV) offers a low-cost and flexible biometric solution to person authentication. While the reliability of ASV systems is now considered sufficient to support mass-market adoption, there are concerns that the technology is vulnerable to spoofing, also referred to as presentation attacks. Spoofing refers to an attack whereby a fraudster attempts to manipulate a biometric system by masquerading as another, enrolled person. Acknowledged vulnerabilities include attacks through impersonation, replay, speech synthesis and voice conversion [1].

There are two general strategies to protect ASV systems from spoofing: the first involves the continued pursuit of more robust ASV technology in the general sense; the second, more popular approach centres around the development of new spoofing countermeasures. Countermeasures have been reported for replay attacks [2, 3, 4, 5], speech synthesis [6, 7, 8, 9], voice conversion [10, 11, 12, 13] and non-speech, artificial signals [14]. For a recent survey, the reader is referred to [1]. While there are currently no alternatives, the use of non-standard databases, protocols and metrics gives rise to two significant problems: (i) a lack of support for comparable and reproducible research, and (ii) countermeasures which lack generalisation.

The focus on highly specific spoofing attacks and the use of non-standard databases often impedes the comparison of different results. For example, much of the work involving voice conversion spoofing attacks was performed with NIST Speaker Recognition Evaluation (SRE) datasets, often with different voice conversion algorithms, protocols and metrics. The Wall Street Journal (WSJ) datasets have been popular in work involving synthetic speech spoofing attacks, but again with a variety of experimental configurations. As a result of database, protocol and metric diversity [15], comparisons between different experimental results is extremely complicated, if not close to meaningless.

Countermeasures which lack generalisation result from the inappropriate use of prior information in their development. The majority of existing countermeasures are optimised with training data produced using exactly the same spoofing method that is to be detected. This is clearly unrepresentative of the real use case scenario in which it is impossible to know the exact nature of a spoofing attack. At best, research results generated with such methodologies exaggerate countermeasure performance; at worst, they mask the true scale of the problem. Generalised countermeasures [16, 17] are needed to detect previously unseen spoofing attacks, i.e. spoofing attacks in the wild.

The ASVspoof challenge aims to encourage further progress through (i) the collection and distribution of a standard dataset with varying spoofing attacks implemented with multiple, diverse algorithms and (ii) a series of competitive evaluations. Following on from the special session in Spoofing and Countermeasures for Automatic Speaker Verification [18] held during the 2013 edition of INTERSPEECH in Lyon, France, the first ASVspoof challenge [19][1] will be held during the 2015 edition of INTERSPEECH in Dresden, Germany. The challenge has been designed to support, for the first time, independent assessments of vulnerabilities to spoofing and of countermeasure performance. The initiative provides a level playing field to facilitate the comparison of different spoofing countermeasures on a common dataset, with standard protocols and metrics. While preventing as much as possible the inappropriate use of prior knowledge, the challenge also aims to stimulate the development of generalised countermeasures with potential to detect varying and unforeseen spoofing attacks.

In order to lower the cost of entry and to maximise participation, the first ASVspoof challenge involved only the detection of spoofed speech. By decoupling spoofing detection from ASV, expertise in the latter was not a prerequisite to participation. Participants were invited to develop spoofing detection algorithms and to submit scores for a freely available, standard dataset and protocol. The dataset was generated according to a diverse mix of 10 different speech synthesis and voice con-

[1] http://www.spoofingchallenge.org

Table 1: Number of non-overlapping target speakers and utterances in the training, development and evaluation sets. The duration of each utterance is in the order of one to two seconds.

Subset	#Speakers		#Utterances	
	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

version spoofing algorithms. The particular spoofing algorithm involved in any trial was not disclosed during the evaluation. Performance was assessed by the organisers using a standard metric described in the evaluation plan [19].

This paper describes the ASVspoof database, protocol and metrics, all of which are now in the public domain. Also presented is a summary of 16 sets of participant results. Finally, observations and findings are presented with priorities for the future.

2. ASVspoof database and protocols

ASVspoof is based upon a standard database consisting of both genuine and spoofed speech[2]. Genuine speech is recorded from 106 human speakers (45 male and 61 female) without any modification, and without significant channel or background noise effects. Spoofed speech is modified from the original genuine speech data by using a number of speech synthesis (SS) and voice conversion (VC) algorithms. More details and protocols to generate the spoofed speech can be found in [20]. The full dataset is partitioned into three subsets, the first set for training, the second for development and the third for evaluation. The number of speakers and trials in each subset is illustrated in Table 1. There is no speaker overlap across the three subsets.

2.1. Training data

The training set includes 3750 genuine and 12625 spoofed utterances collected from 25 speakers (10 male, 25 female). As illustrated in Table 2, each spoofed utterance is generated by one of the five spoofing algorithms (S1 – S5) as follows:

- **S1** - a simplified frame selection (FS) [21, 22] based voice conversion algorithm, in which the converted speech is generated by selecting target speech frames;
- **S2** - the simplest voice conversion algorithm [23] which adjusts only the first mel-cepstral coeffcient (C1) in order to shift the slope of the source spectrum to the target;
- **S3** - a speech synthesis algorithm implemented with the hidden Markov model based speech synthesis system (HTS[3]) using speaker adaptation techniques [24] and only 20 adaptation utterances;
- **S4** - the same algorithm as S3, but using 40 adaptation utterances, and
- **S5** - a voice conversion algorithm implemented with the voice conversion toolkit and with the Festvox system[4].

[2] http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[3] http://hts.sp.nitech.ac.jp/

[4] http://www.festvox.org/

Table 2: Summary of spoofing algorithms implemented in the challenge database. S1 to S5 are known attacks, examples of which are available for system development. S6 to S10 are unknown attacks seen only in the evaluation set. Dev=Development; Eva=Evaluation.

Subset	#trials or #utterances			Vocoder	Spoofing algorithm
	Train	Dev	Eva		
Genuine	3750	3497	9404	None	None
S1	2525	9975	18400	STRAIGHT	VC
S2	2525	9975	18400	STRAIGHT	VC
S3	2525	9975	18400	STRAIGHT	SS
S4	2525	9975	18400	STRAIGHT	SS
S5	2525	9975	18400	MLSA	VC
S6	0	0	18400	STRAIGHT	VC
S7	0	0	18400	STRAIGHT	VC
S8	0	0	18400	STRAIGHT	VC
S9	0	0	18400	STRAIGHT	VC
S10	0	0	18400	None	SS

These five algorithms were chosen since they are among the most easily implemented. They are referred to as *known attacks*, examples of which are available for the training of spoofing detectors. S1 and S2 are two of the most easily implemented VC techniques. S3, S4 and S5 are all implemented with open-source toolkits.

For S1, S2, S3 and S5, 20 utterances were used to train the VC and SS algorithms. These utterances were included in the larger adaptation set used to generate S4. As illustrated in Table 2, S1, S2, S3, and S4, all use the same STRAIGHT vocoder [25] for synthesis, whereas S5 uses an MLSA vocoder [23].

2.2. Development data

The development dataset includes both genuine and spoofed speech from a subset of 35 speakers (15 male, 20 female). There are 3497 genuine and 49875 spoofed trials. Spoofed speech is generated according to one of the same five spoofing algorithms used to generate the training dataset. All data in the development dataset may be used for the design and optimisation of spoofing detectors/countermeasures, for example, to tune classifier hyper-parameters. Spoofing algorithms used to create the development dataset are a subset of those used to generate the evaluation dataset. The aim is therefore to develop a countermeasure which generalises well to spoofed data generated with different spoofing algorithms.

All meta information, including speaker identities and exact spoofing algorithms was provided to challenge participants for both training and development sets. Participants were allowed to use this information for system optimisation.

2.3. Evaluation data

The evaluation set is comprised of 9404 genuine and 184000 spoofed utterances collected from 46 speakers (20 male, 26 female). Recording conditions for genuine speech are exactly the same as those for the training and development sets. However, spoofed data are generated according to more diverse spoofing algorithms. They include the same five algorithms used to generate the training and development sets and an additional five spoofing algorithms, all referred to as *unknown attacks*:

- **S6** - a VC algorithm based on joint density Gaussian mixture models and maximum likelihood parameter generation considering global variance [26];
- **S7** - a VC algorithm similar to S6, but using line spectrum pair (LSP) rather than mel-cepstral coeffecients for spectrum representation;
- **S8** - a tensor-based approach to VC [27] for which a Japanese dataset was used to construct the speaker space;
- **S9** - a VC algorithm which uses kernel-based partial least square (KPLS) to implement a non-linear transformation function [28] (without dynamic information, for simplification), and
- **S10** - an SS algorithm implemented with the open-source MARY Text-To-Speech system (MaryTTS)[5].

S6, S7, S8 and S9 all use 20 utterances to train the conversion function. This is the same training data used for S1, S2, S3 and S5. The speech synthesis system of S10 is trained with 40 utterances per speaker. Since the evaluation set contains spoofing attacks not seen in the development set, it is more representative of the practice scenario (in which there is always the potential for previously unseen attacks). Corresponding results will therefore shed light into the potential for countermeasures 'in the wild', i.e. performance in the face of previously unseen attacks. Participants were requested to submit spoofing detection scores on this set for which **no** meta information was provided.

3. Motivation: degraded ASV performance under spoofing attacks

In order to confirm vulnerabilities to spoofing, experiments were conducted using the challenge database and a state-of-the-art Probabilistic Linear Discriminant Analysis (PLDA) [29, 30] ASV system. Five utterances from each target speaker were used as enrolment data. The Wall Street Journal (WSJ0, WSJ1 and WSJCAM) and Resource Management (RM1) databases were used to train the Universal Background Model (UBM) and eigenspaces. More details of the PLDA system can be found in [20].

ASV results for the evaluation set are presented in Table 3. Results are illustrated for the baseline and for the same system when subjected separately to each of the 10 spoofing attacks. The baseline Equal Error Rate (EER) is 0.42%; the database is clean without any channel or noise effects. Performance degrades significantly when subjected to each spoofing attack. The lowest EER is 0.87% (S2). The highest is 45.79% (S10). These results confirm vulnerabilities to spoofing and demonstrate the importance of developing countermeasures.

4. Protocol, metric and results

4.1. Protocol

ASVspoof 2015 focuses on a stand-alone spoofing detection task. The challenge database is accompanied with a standard protocol. It comprises a list of trials, each corresponding to a randomly named audio file. Participants should assign to each trial a real-valued, finite score which reflects the relative strength of two competing hypotheses, namely that the trial is genuine or spoofed speech. For compatibility with NIST speaker recognition evaluations, we assume that the positive class represents the 'non-hostile' class, i.e. genuine speech. High detection scores are thus assumed to indicate genuine speech whereas low scores are assumed to indicate spoofed speech.

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

Table 3: PLDA ASV system performance. Results illustrated for the baseline and the same system when subjected to spoofing (S1-S10). EER=Equal Error Rate.

Spoofing algorithm	EER (%)
Baseline	0.42
S1	32.92
S2	0.87
S3	25.42
S4	28.44
S5	35.92
S6	33.76
S7	29.71
S8	30.63
S9	29.50
S10	**45.79**
Average(S1-S10)	29.30

4.2. Metric

Participants were not required to optimise a decision threshold, and thus neither to produce hard decisions; the primary metric for ASVspoof 2015 is the 'threshold-free' EER. For the spoofing detection task the EER is defined as follows. Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ denote the false alarm and miss rates at threshold θ:

$$P_{\text{fa}}(\theta) = \frac{\#\{\text{spoofed trials with score} > \theta\}}{\#\{\text{total spoofed trials}\}},$$

$$P_{\text{miss}}(\theta) = \frac{\#\{\text{genuine trials with score} \leq \theta\}}{\#\{\text{total genuine trials}\}},$$

so that $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ are, respectively, monotonically decreasing and increasing functions of θ. The EER corresponds to the threshold θ_{EER} at which the two detection error rates are equal i.e. $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. EERs were estimated using the Bosaris toolkit[6]. While the EER was determined independently for each spoofing algorithm, the average EER for the full evaluation dataset was used for ranking submission results.

4.3. Results

Participants were able to submit scores for up to six systems. One of these systems was designated as the *primary submission*. Spoofing detectors for all primary submissions were trained using only the training data in the ASVspoof 2015 corpus. The dataset was requested by 28 teams from 16 countries; 16 teams returned primary submissions by the deadline. A total of 27 additional submissions were also received. Anonymous results were subsequently returned to each team who were then invited to submit their work to the ASVspoof special session for INTERSPEECH 2015.

This paper summarises the challenge results for primary submissions only. EER results are illustrated in Table 4 in which each line represents the submission of each team. Results are shown independently for known attacks (S1-S5), unknown

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

Table 4: Summary of primary submission results in the ASVspoof 2015 challenge.

System ID	Equal Error Rates (EERs)		
	Known attacks	Unknown attacks	Average
A	0.408	**2.013**	**1.211**
B	0.008	3.922	1.965
C	0.058	4.998	2.528
D	**0.003**	5.231	2.617
E	0.041	5.347	2.694
F	0.358	6.078	3.218
G	0.405	6.247	3.326
H	0.670	6.041	3.355
I	0.005	7.447	3.726
J	0.025	8.168	4.097
K	0.210	8.883	4.547
L	0.412	13.026	6.719
M	8.528	20.253	14.391
N	7.874	21.262	14.568
O	17.723	19.929	18.826
P	21.206	21.831	21.518
Average	3.337 (STD: 6.782)	9.294 (STD: 6.861)	6.316 (STD: 6.558)

attacks (S6-S10) and the average, ranked according to the latter. Almost all submissions achieved excellent performance for known attacks (for which training data was provided). EERs in the case of unknown attacks are significantly and universally higher. The lowest EER for all attacks is 1.211%, whereas those for known and unknown attacks are 0.003% and 2.013%, respectively. The lowest EER for unknown attacks (2.013%) is 671 times higher than that for known attacks (0.003%).

These results illustrate the potential of over-fitting countermeasures to known attacks which may leave ASV systems prone to unforeseen spoofing attacks. For example, while system **D** achieves a lower EER than system **A** in the case of known attacks (0.003% vs 0.408%), the EER for system **D** is over twice that of system **A** in the case of unknown attacks (5.231% vs 2.013%). In turn these results thus confirm the importance of developing more generalised countermeasures and also the need for further work and future evaluations.

5. Discussion and future work

Discussed here are some of the limitations of the ASVspoof challenge and priorities for future research. One limitation regards the inclusion of only high-technology speech synthesis and voice conversion spoofing algorithms. While their relative severity is currently uncertain, low-technology replay and impersonation attacks were not considered. Even if these alternative attacks prove to be less severe than speech synthesis and voice conversion, they might well be the most common in practice; their implementation requires no particular expertise, nor equipment. There is thus some cause to include such attacks in future ASVspoof challenges.

The second limitation relates to the focus on STRAIGHT vocoders. Other types of vocoder, such as sinusoidal vocoders [31] are also popular and their use may have different impacts on spoofing. Accordingly, a greater variety of vocoders and potentially more advanced spoofing algorithms should be considered in future challenges.

The lack of any additive noise or channel effects may also be a limitation. Even if their omission for the first evaluation was a deliberate choice, their effect on spoofing and spoofing detection is currently uncertain. It will thus be important to address additive noise and channel variability in the future.

Future evaluations should also measure the impact of spoofing and detection on ASV. While such work has already been reported, in many cases it considered spoofing attacks implemented with full knowledge of the ASV system. Future evaluations should thus address the integration issue.

It is also stressed that the evaluation was not intended, nor sufficient to compare the relative severity of different spoofing attacks; different levels of effort have been dedicated to developing speech synthesis and voice conversion attacks and different quantities of training data were used in their implementations. A meaningful comparison between the severity of each attack should be a priority for the future. In addition, closer collaboration with the speech synthesis and voice conversion communities should be considered in order that future evaluations include the very best algorithms.

Finally, the focus on text-independent ASV is perhaps not the most representative of authentication applications in which spoofing is relevant. Future evaluations should therefore include an emphasis on text-dependent ASV. The organisers are currently working in this direction.

6. Conclusions

The first automatic speaker verification spoofing and countermeasures challenge (ASVspoof 2015) was highly successful in attracting significant participation. This paper presents the challenge database, organisation, evaluation results, and priorities for future challenges and research.

Even though the best results show an overall average detection EER of less than 1.5%, the EER of unknown attacks is five times greater than that of known attacks. In addition, while some attacks are easily and consistently detected, others (e.g. S10) provoke extremely high error rates nearing 50%. The low overall average is not necessarily the best indicator of ASV robustness especially if, as is likely, fraudsters concentrate on the most successful spoofing attacks. Accordingly, the error rates for the most effective spoofing attacks are perhaps more representative than the average error rates. In any case, even if average detection EERs are low, it is the resulting degradation in automatic speaker verification performance which is of the greatest importance; these degradations are often much greater.

The overriding findings from ASVspoof 2015 show the remaining need to develop more generalised spoofing detection algorithms. Generalisation will remain a focus for future evaluations, as will the integration of spoofing detection with automatic speaker verification and text-dependent scenarios.

7. Acknowledgements

We would like to thank Dr. Daisuke Saito from University of Tokyo, Prof. Tomoki Toda from Nara Institute of Science and Technology, Mr. Ali Khodabakhsh and Dr. Cenk Demiroglu from Ozyegin University, and Prof. Zhen-Hua Ling from University of Science and Technology of China for their contributions to the spoofing materials used in the challenge. The work was partially supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology), Academy of Finland (project no. 253120 and 283256) and by the TABULA RASA project funded under the 7th Framework Programme of the European Union (EU) (grant agreement number 257289).

8. References

[1] 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, no. 0, pp. 130 – 153, 2015.

[2] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *IEEE Int. Carnahan Conf. on Security Technology (ICCST)*, 2011.

[3] ——, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, ser. Lecture Notes in Computer Science, C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, Eds. Springer, 2011, pp. 274–285.

[4] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Proc. Int. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2014.

[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, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[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] 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.

[9] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "A cross-vocoder study of speaker independent synthetic speech detection using phase information," in *Proc. Interspeech*, 2014.

[10] 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.

[11] 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.

[12] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *Proc. Interspeech*, 2013.

[13] 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.

[14] F. Alegre, R. Vipperla, N. Evans *et al.*, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *Proc. Interspeech*, 2012.

[15] N. Evans, J. Yamagishi, and T. Kinnunen, "Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics," *IEEE Signal Processing Society Speech and Language Technical Committee Newsletter*, 2013.

[16] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[17] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and anti-spoofing in the i-vector space," *IEEE Trans. on Information Forensics and Security (to appear)*, 2015.

[18] N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013.

[19] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *http://www.spoofingchallenge.org/asvSpoof.pdf*, 2014.

[20] 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.

[21] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, "Towards a voice conversion system based on frame selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

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

[23] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992.

[24] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained smaplr adaptation algorithm," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[25] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[26] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[27] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. Interspeech*, 2011.

[28] 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.

[29] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, 2010.

[30] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince, "Probabilistic models for inference about identity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012.

[31] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.