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

Recently, an increasing number of 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 competitive 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 in [2, 3, 4, 5], for speech synthesis in [6, 7, 8, 9], voice conversion in [10, 11, 12, 13] and non-speech, artificial signals [14]. For a recent survey, the reader is pointed 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 lack of support for comparable research stems from the focus on specific spoofing attacks and the use of non-standard databases. 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 learned using training data produced using exactly the same spoofing method that is to be detected. Hence, the countermeasure is designed with full knowledge of the spoofing attack algorithm. 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 held during the 2013 edition of INTERSPEECH in Lyon, France [18], the first edition of ASVspoof challenge[1] will be held at the 2015 edition of INTERSPEECH in Dresden, Germany [19]. 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 standard dataset, with standard protocols and metrics. While preventing as much as possible the inappropriate use of prior knowledge, and 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 attract the widest possible participation, the first ASVspoof challenge related only to 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,

[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

Table 2: Summary of spoofing algorithms implemented in the challenge database. S1 to S5 are known attacks seen at system development stage while S6 to S10 are novel attacks seen only at 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 [21]	VC
S2	2525	9975	18400	STRAIGHT	VC
S3	2525	9975	18400	STRAIGHT	SS
S4	2525	9975	18400	STRAIGHT	SS
S5	2525	9975	18400	MLSA [22]	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

standard dataset and protocol. The dataset is generated according to a diverse mix of 10 different speech synthesis and voice conversion spoofing algorithms. The particular spoofing algorithm involved in any trial is not disclosed during the evaluation in order to encourage the development of generalised countermeasures. 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 anonymous summary of 16 sets of participant results. Finally, observations and findings are presented with directions 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) spoofing techniques. 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 regarding target speakers used in voice conversion and speech synthesis adaptation; this is to encourage the development of speaker-independent countermeasures.

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

2.1. Training data

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

- **S1**: This is a simplified frame selection (FS) [23, 24] based voice conversion algorithm, in which the converted speech is generated by selecting target speech frames.
- **S2**: The algorithm only shifts the first coefficient (C1) of mel-cepstral coeffcients [22] that is to shift the slope of a spectrum, and is the simplest voice conversion algorithm.
- **S3**: This algorithm is implemented by the hidden Markov model based speech synthesis system (HTS[3]) using speaker adaptation techniques [25], and only 20 utterances were used for speaker adaptation.
- **S4**: This is generated by using the same technique as S3, but using more utterances, in particular 40 utterances to do adaptation.
- **S5**: This spoofing algorithm is implemented by the voice conversion toolkit with Festvox system[4].

In S1, S2, S3 and S5, 20 utterances were used for training. These utterances were included in the larger adaptation set for S4. The spoofing algorithms in this set are defined as *known attacks*, assuming some spoofed materials are available to train spoofing countermeasures. The reason to choose these algorithms as known attacks is that they can be easily implemented to train and tune detectors. In particular, S1 and S2 are two of the most easily implemented voice conversion techniques, S3 and S4 are implemented by the opensource hidden Markov model based speech synthesis system (HTS)[5], and S5 is implemented by a publicly available voice conversion toolkit within the Festvox system[6]. In S1, S2, S3, and S4, the same STRAIGHT vocoder is adopted for synthesis. All data in the training set may be used to train spoofing countermeasures.

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 hyper-parameters for classifiers. We note that the 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 has potential to generalise well to spoofed data generated with different spoofing algorithms.

In the challenge, we provide all the meta information, including speaker identies, and exact spoofing algorithms, for both training and development sets. The participants are allowed to use these information to optimise their systems.

[3] http://hts.sp.nitech.ac.jp/
[4] http://www.festvox.org/
[5] http://hts.sp.nitech.ac.jp/
[6] http://www.festvox.org

2.3. Evaluation data

In the evaluation set, there are 9404 genuine and 184000 spoofed utterances collected from 46 speakers (20 male and 26 female speakers). The recording conditions of genuine speech are exactly the same as those for the training and development. However, the spoofed data are generated according to more diverse spoofing algorithms. They include the same five algorithms used to generated the training and development dataset, and another five spoofing algorithms, which are designated as *unknown attacks*. The algorithms to implemented unknown attacks are detailed as follows:

- **S6**: This is a voice conversion algorithm using joint density Gaussian mixture model and maximum likelihood parameter generation considering global variance [26].
- **S7**: This algorithm is similar to S6, but using line spectrum pair (LSP) rather than mel-cepstral coeffecients to represent spectra.
- **S8**: This is a tensor-based voice conversion (TVC) [27], and a Japanese dataset was used to construct the speaker space before implementation conversion function using the English conversion training data.
- **S9**: This voice conversion algorithm uses kernel-based partial least square (KPLS) to implement a non-linear transformation function [28]. Here dynamic information is not used for simplification.
- **S10**: This speech synthesis algorithm is implemented by the opensource MARY Text-To-Speech system (MaryTTS)[7] with 40 utterances from each target speaker as training data.

In S6, S7, S8 and S9, 20 utterances are used as conversion function training data, which are the same as that used for S1, S2, S3 and S5.

Being intentionally different, we could use this setting to simulate practice scenarios, and get some insight into countermeasure performance 'in the wild', i.e. performance in the face of previously unseen attacks (although we have used existing well-known techniques to implement the unseen attacks). All the participants are requested to submit spoofing detection scores on this set, and they do not have access to any meta information, such as gender, speaker ID and spoofing algorithm label, before they submitting scores.

3. Motivation: degraded ASV error rates under spoofing attacks

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

Results are presented for the evaluation set in Table 3. The baseline Equal Error Rate (EER) is 0.42%; the database is clean without any channel or noise effects. Under spoofing, the performance of the PLDA system is degraded significantly by all spoofing algorithms. The lowest EER is 0.87% (S2) and the highest is 45.79% (S10). These results confirm vulnerabilities to spoofing and demonstrate the importance of developing countermeasures.

Table 3: Spoofing performance of the challenge database on a PLDA ASV system. EER=Equal Error Rate.

Spoofing algorithm	EER (%)
No-spoofing	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. Protocols, metrics and results

ASVspoof 2015 focuses on a stand-alone spoofing detection task. The challenge database is accompanied with a standard protocol. They comprise a list of trials, each corresponding to a randomly named audio file of either genuine or spoofed speech. 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.

4.1. Evaluation metric

In the challenge, participants are not required to optimise a decision threshold, and thus neither produce hard decisions; the primary metric for ASVspoof 2015 is the 'threshold-free' *equal error rate* (EER), defined as follows. Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ denote the false alarm and miss rates at threshold θ:

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

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. $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$. In practice, the organisers use the Bosaris toolkit[8] to compute the EERs. While EERs will be determined independently for each spoofing algorithm, the average EER for the full evaluation dataset will be used for ranking.

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

[8] 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	Overall
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)

4.2. Challenge results

In the ASVspoof 2015 challenge, each team was allowed to submit score files from up to six systems, but only one system was used as the *primary submission*, in which the participants can only use the training data provided by organisers to train their detector. There were 28 teams from 16 countries who expressed their interests and requested the challenge database. Among them, 16 teams submitted their primary submissions by the deadline, and they also submitted another 27 optional submissions. In total, the organisers received 43 submissions. The organisers returned the results to each team individually, without disclosing other teams' names.

In this paper, we only summarise the results of primary submissions. The results are presented in Table 4. Each alphabet represents a team without disclosing the participant identities. From the results, it is obvious that most of the teams achieve good performance on known attacks in terms of low EERs. The lowest EER for all attacks is 1.211%, and that for known and unknown attacks are 0.003% and 2.013%, respectively. The EERs for unknown attacks are considerably higher than that for known attacks. In particular, the lowest EER for unknown attacks (2.013%) is 671 times higher than that for known attacks (0.003%).

The results also indicate that even a countermeasure is able to give good performance on known attacks, it may not work that well on unknown attacks. For example, system **D** achieves much lower EER than that of system **A**, 0.003% vs 0.408%, for known attack, however, the EER of system **D** for unknown attacks is 2.656 times higher than that of system **A**. All these observations suggest that it is necessary to develop more generalised countermeasures which are able to achieve robust performance on both known and unknown attacks. It also suggests more efforts should be made in order to resolve the spoofing attack issue.

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 speech synthesis and voice conversion spoofing algorithms. These two, relatively high-technology approaches are not the only attack vectors. Replay and impersonation attacks were not considered, even if the relative severity of such attacks is currently uncertain. 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 special expertise, nor equipment. Their consideration should thus be a priority for future ASVspoof challenges.

The second limitation regards the focus on STRAIGHT vocoders speech signal reconstruction and public softwares. 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 more advance 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 has considered spoofing attacks implemented with full knowledge of the ASV system. Future evaluations should thus address the integration issue while maintaining independence between spoofing attacks and ASV systems.

It is also stressed that the evaluation was not intended, nor sufficient to compare the relative severity of different spoofing attacks; Different levels of time and effort have been dedicated to developing speech synthesis and voice conversion attacks. Closer collaboration with the speech synthesis and voice conversion communities should be considered in order that future evaluations include the very best spoofing 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

This paper presents the first automatic speaker verification spoofing and countermeasures challenge (ASVspoof), evaluation results and directions for future challenges and research. Even the best results show that a baseline detection EER of 0.408% increases almost five-fold when known spoofing attacks are replaced with those unknown to the system. It is stressed that, even if the best EER for unknown attacks of 2.013% may seem low, a small increase in the detection EER can lead to much more significant degradations in automatic speaker verification performance. In this sense the need to develop more generalised spoofing detection algorithms remains. Generalisation will remain a focus for future evaluations, as will the additional consideration of integrated spoofing detection and automatic speaker verification, including text-dependent scenarios.

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.

7. 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] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[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.