Spoofing Attacks to I-vector Based Voice Verification Systems Using Statistical Speech Synthesis with Additive Noise and Countermeasure

Mustafa Caner Özbay, Ali Khodabakhsh, Amir Mohammadi, Cenk Demiroğlu
Electrical and Computer Engineering Department
Özyeğin University, 34794, Istanbul, Turkey
Email: {mustafa.ozbay}@ozu.edu.tr, ali.khodabakhsh, amir.mohammadi, cenk.demiroglu@ozyegin.edu.tr

***Abstract*—Even though improvements in the speaker verification (SV) technology with i-vectors increased their real-life deployment, their vulnerability to spoofing attacks is a major concern. Here, we investigated the effectiveness of spoofing attacks with statistical speech synthesis systems using limited amount of adaptation data and additive noise. Experiment results show that effective spoofing is possible using limited adaptation data. Moreover, the attacks get substantially more effective when noise is intentionally added to synthetic speech. Training the SV system with matched noise conditions does not alleviate the problem. We propose a synthetic speech detector (SSD) that uses session differences in i-vectors for counterspoofing. The proposed SSD had less than 0.5% total error rate in most cases for the matched noise conditions. For the mismatched noise conditions, missed detection rate further decreased but total error increased which indicates that some calibration is needed for mismatched noise conditions.**

***Index Terms*—spoofing attacks, speaker verification, statistical speech synthesis, speaker adaptation, synthetic speech detection**

I. Introduction

There has been substantial progress in the speaker verification (SV) field in recent years [1]. I-vector based approach in particular received significant attention due to its high performance. However, despite the success of the i-vector method in verification, it has been shown to be vulnerable to spoofing attacks [2], [3], [4]. Some of the prior methods for spoofing the SV systems and detection of spoofing attacks are described below.

In [5], GMM-based voice transformation is proposed using parallel data. To increase the effectiveness of the attacks, segments of speech that get high scores from the voice verification system are repeated. Two countermeasures are also proposed in [5]. In one approach, distributions of Gaussian components are used to detect repetitions of Gaussians in speech. In a second approach, automatic voice quality assessment tools are used to detect synthetic speech.

If a speech vocoder is used during an attack, phase spectrum can be used to detect the synthetic speech as proposed in [6]. However, in many speech applications, only the spectral magnitude features are transmitted to avoid increasing the network traffic and minimize the delay. Our focus here is detection of attacks when only the Mel-frequency Cepstral coefficients (MFCCs) are available at the detector.

Modified speech detection performance when the detector is trained with different kind of voice conversion techniques is reported in [6]. Modulation of spectral features over longer durations is investigated in [7]. Longer duration modulation features were found to be complementary to short-time features.

Voice conversion methods typically require significant amount of parallel data to be successful. However, in many practical cases, the attacker is required to attack the verification system with very limited amount of adaptation data to be able to spoof a large number of accounts. Statistical speech synthesis (SSS) systems are particularly suitable for such attacks since adaptation with a couple of utterances are feasible in those systems [8], [9], [10]. Therefore, we focused on the SSS systems here. Experiment results show that effective spoofing is possible with only a couple of utterances in clean training and test conditions.

Even though SSS is effective at spoofing, synthetic speech with SSS can be detected by exploiting its overly smooth nature [11]. Here, we investigated the possibility of attacking the system by intentionally adding noise to synthetic speech with the hypothesis that noise can reduce the smoothness of synthetic speech and make it more difficult to detect. Noises at and above 10dB are added to synthetic speech because utterances at those signal to noise ratio (SNR) values are expected to be common in real-life. We have found that the attacks get substantially more effective when noise is added to synthetic speech even when the verification system is trained with matched noise conditions.

Besides showing the effectiveness of the method for attack, we propose a novel and simple synthetic speech detector that uses session differences in i-vectors to detect between synthetic speech. We then experimentally show that the proposed detector has error rates less than 0.5% in all test conditions. To make the problem more challenging, we used more advanced techniques such as global variance (GV) [12] and STRAIGHT vocoding [13] on the attacker side but not on the detection side. Even when there is such mismatch between training and test data, the detector is found to perform well in most cases.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

II. Front-end factor analysis (FA)

Gaussian Mixture Models (GMM) are typically used to represent the acoustic feature space in speaker verification systems. In most of the current systems, a universal background model (UBM) is first trained and then speaker-specific models are obtained by adapting the UBM using a Maximum A Posteriori adaptation (MAP) approach.

Typically, supervector of mean vectors in UBM is very high dimensional which increases the number of parameters to adapt. In the factor analysis (FA) approach [14], mean vectors of speakers, m_s, are represented in a lower dimensional total variability space in which

$$m_s = m_0 + Tw_s \tag{1}$$

where w_s is called an identity vector (i-vector). T matrix is trained using a database where multiple sessions are available for each speaker.

In enrollment, an i-vector is extracted from each enrollment speaker. In testing, the i-vector is extracted from the test utterance and compared with the i-vector of the target speaker. Probabilistic linear discriminant analysis (PLDA) [15] is used for scoring here.

III. Detection of synthetic speech

Even though removing the session effects from the i-vectors is important for successful verification, session differences contain valuable information for detecting synthetic speech. For session-i, channel vector can be defined as

$$m_{c,i} = m_{s,i} - m_s \tag{2}$$

where $m_{s,i}$ is the i-vector extracted in session-i and m_s is the mean i-vector for speaker s.

Channel vectors contain information about the distortions that are session-specific. In the case of synthetic speech, there is additional variability. For example, it is well-known that synthetic features are smoother than natural features which reduce the variance of all features [11]. Moreover, because feature vectors in close proximity are similar to each other, they are assigned to the same Gaussian. Therefore, as opposed to the variety of Gaussians in natural speech, fewer Gaussians are observed with higher frequency in synthetic speech.

We investigated the differences between i-vectors of synthetic and natural speech through visualization. To that end, Fisher linear discriminant analysis (LDA) is used to reduce dimensionality of the channel vectors to 2. Channel vectors of synthetic and natural speech is compared in Fig. 1. In the clean case, there is a clear separation between synthetic and natural vectors. In the noisy case, the two clusters are still clearly separable. However, the margin is not as large as the clean case. Thus, noise distorts the smooth structure of the synthetic features and make clean and noisy channel less separable.

Even though the clusters are separable in noisy conditions, an important question arises: what if the attacker and the defender use different SSS technologies? In particular, we are interested in the worst case where the attacker has more advanced technology compared to the defender. To test that condition, STRAIGHT vocoding and GV adjustment is used at the attacker side but not at the defender side. Clusters for synthetic and natural channel vectors at 10dB babble noise are shown in Fig 1. Using different synthesis technologies by the attacker and defender caused significant overlap between the clusters which makes the detection problem harder.

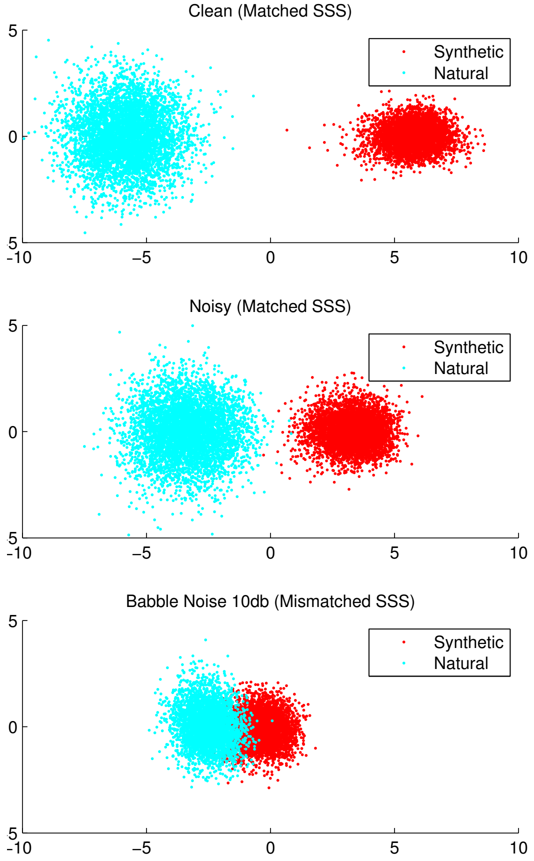

Fig. 1. Illustration of channel vectors after they are mapped to 2 dimensions using LDA. In the top figure, clean synthetic and natural data is used where both test and train synthetic data are generated with STRAIGHT and GV. In the middle figure, noisy natural and synthetic data are used where both test and train synthetic data are generated with STRAIGHT and GV. Mixed type of noises are used in training LDA and channel vectors of noisy natural and synthetic speech (mixed noise) are shown. In the bottom figure, LDA is trained on noisy synthetic speech without GV and STRAIGHT but the test data are generated with STRAIGHT and GV. Effect of mismatch in synthesis technologies are shown. Mixed type of noises are used in training LDA and channel vectors of noisy natural and synthetic speech (babble noise) is shown.

Exploiting the structure in the distribution of channel vectors, a detector is designed to detect synthetic speech. Dimensionality of session vectors are first reduced using LDA. Then, a support vector machine (SVM) with soft-decision output is trained with the noisy synthetic and noisy natural session vectors. Linear kernel is used with the SVM.

IV. Experiments

A. Experimental Setup

WSJ1 database [16] is used for the verification experiments similar to [2]. 69 male test speakers are enrolled into the system. Each enrollment utterance is around 4-6 seconds long. For each enrolled speaker, 59 client tests and 340 impostor tests are done. Impostor tests are created by using 5 utterances from each of the 68 impostor speakers among the enrolled speakers. Each test is done using one utterance. Verification system uses 19 dimension MFCC plus 1 energy static features and their delta and delta-delta features. However, static energy is not used which makes the total dimension of features 59. 256 mixture UBM is trained using 84 male speakers, and 60 utterances from each speaker. T matrix is trained using those same speakers and utterances. Rank of the T matrix is set to 400.

Experiments are done for clean training and test data as well as noisy training and test data. Noise is added to clean speech samples at 10, 15, and 20dB SNRs because when the SNR is below 10 dB, performance of the verification system is found to be unacceptably poor. The detector and the verification systems are trained using a mixture of white, babble, car, and station noisy samples under 10, 15, and 20dB SNRs in noisy conditions. Bus, cafe, metro, and office noises are used only during testing.

For each enrolled speaker, different statistical models are created for attacks using adaptation with one, two, three, and four utterances. Synthesis is done for all of the 69 speakers enrolled into the verification system. Enrollment and test data are not used for adaptation. Experiments are also done using speaker dependent (SD) models for comparison purposes, where 150 utterances are used for adaptation. Speaker-independent (SI) model is generated using four male speakers and 1250 utterances from each speaker. Constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm is used for adaptation [9].

SSS systems were trained with 198 dimensional vectors consisting of 40 Mel-Generalized Cepstral (MGC), 1 Log-Fundamental frequency (LF0), and 25 Band APeriodicity (BAP) coefficients and their delta and delta-delta parameters. 25 msec analysis window with 5 msec frame rate is used for feature extraction. Phonemes are modeled with 5 state hidden semi-Markov models (HSMM) [17]. STRAIGHT vocoding and global variance adjustments are done to improve the synthesis quality.[13]

Training data for UBM and T are used for training the detectors. The same features used in the verification system are used for the detector. Similar to the attacker, a speaker-independent (SI) model is needed for creating the synthetic speech database for training the detector. Here, SI model is trained using the training data of the verification system. Synthesized versions of the test data used for testing the verification system are used to assess the performance of the detectors under different conditions. Detector performance is reported in terms of equal-error-rate (EER) for each test condition. Dimension of the channel vectors are reduced to 50 with LDA before using SVM for synthetic speech detection.

B. Results and discussion

Baseline performance of the voice verification system in clean training and test conditions in terms of equal-error-rate (EER) is 0.23%. Performance of the system for individual noise types and SNRs are shown in Table I. EER calculated under all SNRs and noise types combined is 1.81% which is almost 8-folds increase compared to clean conditions. White noise had particularly higher error rate compared to others since it distorts all of the speech spectrum.

TABLE I
EER OF THE VOICE VERIFICATION SYSTEM FOR DIFFERENT NOISE TYPES AND SNRS. VERIFICATION SYSTEM IS TRAINED WITH MIXED NOISE CONDITIONS AND SNRS. WHITE, BABBLE, CAR, AND STATION NOISES WERE USED IN TRAINING OF THE VERIFICATION SYSTEM.

Seen noises	10db	15db	20db
white	4.53	1.98	1.16
babble	1.27	1.23	1.11
car	1.21	1.19	1.26
station	0.96	0.97	1.03

Unseen noises	10db	15db	20db
bus	1.27	1.24	1.22
metro	1.26	1.10	1.13
office	1.25	1.28	1.25
cafe	1.13	1.13	1.15

For spoofing attacks, threshold of the voice verification system is set to 1.81% average EER point. Results with clean train/test and noisy train/test are shown in Fig. 2. Noise substantially increases the effectiveness of the attacks. Effectiveness of car and bus noises are below others since those noise types have lower bandwidth. Interestingly, effectiveness of the attacks are close to each other at different SNRs. This is thought to be a result of the fact the system is trained with a mix of all SNRs and all noises. Moreover, the calibration is also done with a mix of all conditions. Thus, the system does not seem to substantially favor any particular SNR.

Spoofing attacks become more effective when more adaptation data becomes available. However, performance seems to saturate more rapidly in the clean conditions compared to noisy conditions.

White noise has especially lower false alarm rates compared to other noise types. The reason for that can be understood from Fig. 3. In that figure, at 10db, white noise detection error trade-off (DET) curve is significantly separated from the other noise types which holds for other SNR types and adaptation data sizes also. The 1.81 % EER, however, is computed by using all noise conditions at all SNRs which causes an outlier effect where the white noise has a big effect on the operating point. Thus, at the 1.81 % operating point, all noises other than white noise have significantly higher false alarm rates compared to missed detection rates as shown in Fig. 3. White noise, however, does not significantly deviate from the EER point. As a result, its false alarm rate is lower than others in spoofing attacks.

The proposed detector has 0% detection error for clean case. For noisy case, EER is less than 0.5% for all noise and SNR conditions as shown in Fig. 4. Thus, synthetic speech

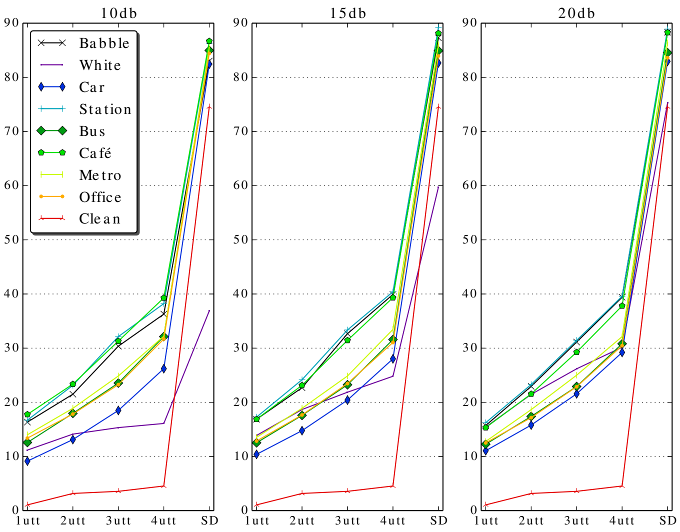

Fig. 2. Verification false alarm rates under attack with synthetic speech. Results are reported for both clean and noisy conditions. In the "Clean" case, both test and train samples are clean and it is shown in the figures for comparison purposes. Babble, cafe, and station noise results have almost overlapped here. Metro, bus, and office noise results have almost overlapped here.

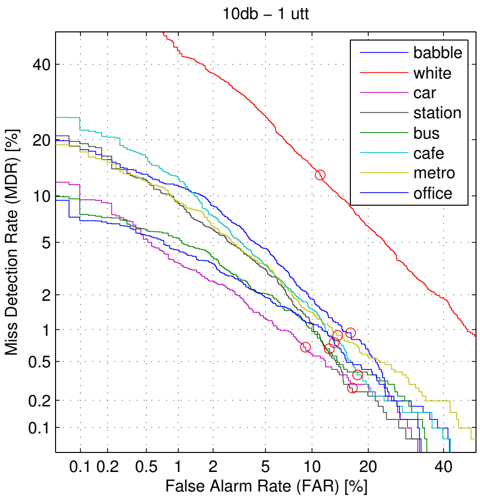

Fig. 3. DET curves of the verification system under attack at different noise conditions at 10dB. Natural speech is used for clients and synthetic speech is used for impostors. Performance of the verification system for different noise types are indicated with circles when the verification system is tuned to 1.81% EER point with mixed noise conditions.

can be effectively detected in the i-vector space with very high accuracy as observed visually in Section 3. To check if these results still hold for mismatched SSS technologies in attacker and defense sides, the detector is trained with SSS without GV or STRAIGHT. The attacker, however, used STRAIGHT and GV which are known to increase the quality of speech. Effectiveness of the spoofing attacks in such mismatch conditions are reported in Fig. 5. Under the mismatched SSS synthesis conditions, detection performance decreases substantially especially for babble and white noises.

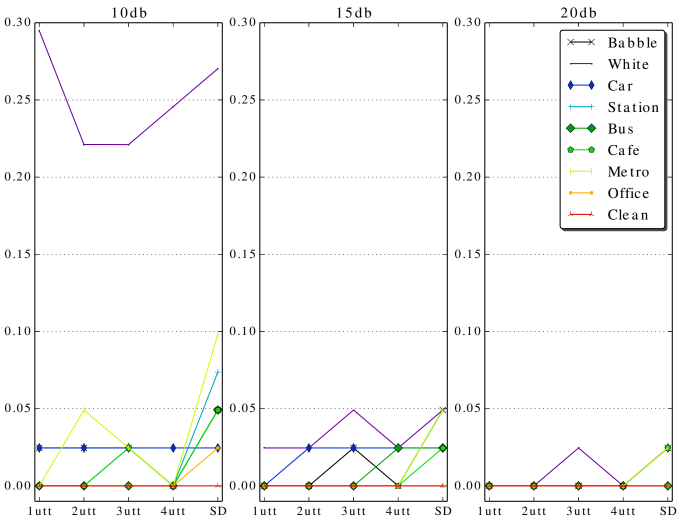

Fig. 4. Detector performance (EER) when detector is trained with STRAIGHT vocoder and GV. And, the attacker uses STRAIGHT vocoder and GV as well (Matched condition in SSS). Except for white noise at 10dB, EER of all cases is under 0.1%

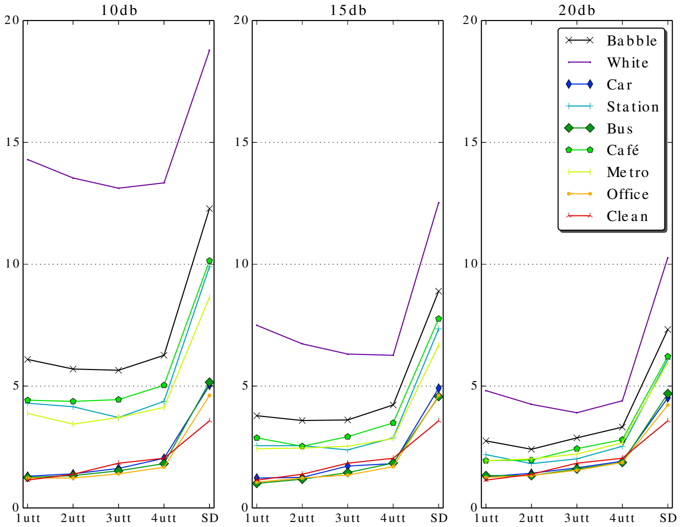

Fig. 5. Detector performance when detector is trained without STRAIGHT vocoder or GV but attacker uses those two techniques for generating more natural speech. Metro, cafe, and station noise results have almost overlapped here. Car, bus, and office noise results have almost overlapped here.

This result calls for training detectors with different synthesis conditions and not fit the detector on one particular type of SSS.

V. Conclusion and Future Work

In this paper, we propose to attack an i-vector based voice verification system with SSS when limited amount of adaptation data is available. We have shown that effective attacks are possible in clean conditions. Moreover, substantial performance gains are obtained when the verification system is trained with mixed noise conditions at and above 10 dB and noise is intentionally added to synthetic speech. We also proposed a synthetic speech detector that is found to have excellent performance in noisy conditions.

The proposed detector did not perform as well when different SSS vocoders are used for training and testing the detector. In the future work, we will focus increasing the robustness of the detector to mismatch in SSS techniques.

Acknowledgment

This work was supported by TUBITAK 1001 grant No 112E160.

References

[1] A. F. Martin, M. Yadagiri, G. R. Doddington, C. S. Greenberg, and V. M. Stanford, "The 2012 nist speaker recognition evaluation," in *NIST SRE 2012 Workshop*, Dec 2012.

[2] 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[3] 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 *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2012, pp. 4401–4404.

[4] N. W. D. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *INTERSPEECH*, Lyon, FRANCE, Aug 2013.

[5] F. Alegre, R. Vipperla, and N. Evans, "Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals," in *INTERSPEECH, 13th Annual Conference of the International Speech Communication Association*, 2012.

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

[7] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *ICASSP*, 2013.

[8] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing. ICASSP, IEEE International Conference on*, vol. 4, 2007.

[9] 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, 2009.

[10] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu *et al.*, "Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 984–1004, 2010.

[11] F. Alegre, R. Vipperla, A. Amehraye, and N. W. D. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, FRANCE, 08 2013.

[12] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[13] 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.

[14] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[15] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[16] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*. Citeseer, 2007, pp. 294–299.