Pitch-shift robustness in voice-based fraud detection

David Looney
Pindrop Ltd. UK
dlooney@pindrop.com

Nikolay D. Gaubitch
Pindrop Ltd. UK
nick@pindrop.com

***Abstract*—Voice recognition is a powerful means to guard against known bad actors in speech-based applications. Pitch-shift software, however, is widely available and can facilitate the disguise of a bad actor's voice from such systems. We here provide insight into the level of risk associated with different kinds of pitch-shift by demonstrating how formant-preserving ones, such as the time-domain pitch-synchronous overlap-add (TD-PSOLA), have considerably lower impact on voice recognition than those that do affect the formants, such as the waveform-similarity overlap-add (WSOLA). Finally we propose a combination of two complimentary methods, augmentation of the speaker models and pitch-change detection using voice recognition features, to prevent pitch-shift-enabled evasion in fraud detection systems.**

***Index Terms*—pitch-shift detection, voice security**

I. Introduction

The performance of voice recognition has increased significantly over the past decade, mostly due to the use of deep neural networks (DNNs) in combination with speaker embeddings [1], commonly referred to as x-vectors [2], for speech representation. This has led to a growing use of voice recognition tools in fields such as forensics and user authentication. Meanwhile, use of the voice channel for malicious purposes has been growing at a fast pace. In the retail banking industry, for example, fraudulent activity, such as account takeover, often involves phone calls from malicious actors. Voice identification, i.e. one-to-many voice recognition systems, with speaker models trained on speech utterances from these actors can be an effective strategy for fraud detection.

However, a common tactic employed by malicious actors to deliberately disguise their identity in the voice channel is pitch-shift. It can be achieved either manually, for example, by straining their voice so as to raise or lower the pitch, or by using software [3]–[5] using one of many free application available for smartphones and personal computers. Despite the recent developments in voice conversion and speech synthesis tools [6], pitch-shift remains a computationally cheap, readily-available tool and an important problem to address in the context of voice identification [4], [7].

Our focus here is on the different forms of software-driven pitch-shift and their effects on voice identification, and mitigating against those effects in fraud detection scenarios. Note the inherent challenges are different to those in voice spoofing and impersonation where the objective is to deceive a voice verification system, that is a one-to-one match with a claimed voice identity. This issue is studied extensively by, for example, the ASVSpoof community [8], [9]. Instead we propose solutions to enhance a voice identification system by (i) a speaker augmentation strategy for small pitch-shifts and (ii) leveraging a dedicated pitch-change detection model for large pitch-shifts. Such a system is shown in Fig. 1. We concede that attacks uniquely detected by a pitch-change model do not enable identification of the actor, but this is acceptable as the system we are considering is primarily concerned with fraud detection and not voice identification.

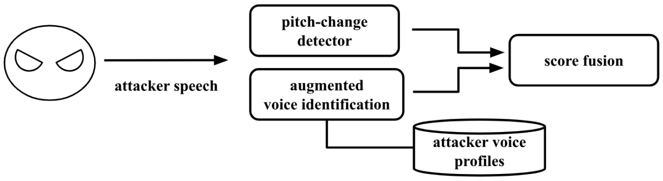

Fig. 1. **Fraud detection system based on voice identification**. The proposed system is robust to small pitch-shift attacks with the assistance of enrolment augmentation and larger ones via dedicated pitch-shift detection.

The main contributions of this work are as follows. In Section III, we demonstrate that formant-preserving pitch-shifting methods such as TD-PSOLA have only minor impact on voice identification compared to non-formant preserving methods such as WSOLA. Thus, the latter is more important in voice identification evasion. In Section IV we show how pitch-shift augmentation when training a speaker model for voice identification is effective for subtle pitch-shift but less so for the severe cases. On the other hand, we show that x-vectors facilitate accurate pitch-shift detection compared with standard speech features, such as the mel frequency cepstral coefficients (MFCCs), and that the detection rate is high for more severe cases of pitch-shift. Finally, we evaluate the pitch-shift invariance of a fraud detection system created from these two complimentary components.

II. Relation to prior work

Pitch-shift detection has previously been linked to the performance of speaker recognition [3], [5], [11]. However, the detection of pitch-shift is typically treated as a standalone problem rather than as part of a fraud detection system. The body of existing work includes various feature representations of the speech signal, including MFCCs [12], linear frequency cepstral coefficients (LFCCs) [13] and features stemming from the the source-filter speech production model [16] such as formant values and the kurtosis of the linear prediction coding (LPC) residual. The features have then been used to train binary classifiers based on Gaussian mixture models (GMMs) or support vector machines (SVMs) [12]–[14], or one-class classifiers [15]. We propose here to use x-vectors as an alternative, promising signal representation with a low-cost DNN classifier for pitch-shift detection.

In [14] the authors consider an end-to-end convolutional neural network (CNN) approach that aims to estimate the amount of pitch shift (in quantised semitones) instead of the detection of pitch-shifted voice. The work in [11] investigates several voice modification techniques and performs a basic pitch-shift inversion in the attempt to improve speaker identification performance in the presence of pitch-shift. In [17] the authors attempt to design pitch-insensitive speaker recognition using sound field analysis. While this is potentially a promising idea, we demonstrate that there is a trade-off between recognition accuracy and pitch-shift robustness. Instead, we introduce the concept of pitch-shift robustness via augmentation in the speaker model training and we combine that with the above-mentioned detector for an overall pitch-shift robust fraud detection system.

Lastly, voice modification detection studies typically categorise different methods for voice pitch-shift by different audio processing software packages [11], [15], or by considering time-domain versus frequency-domain implementations [14]. While this categorisation may be relevant for some cases of pitch-shift detection it is not as relevant when discussing the effects of pitch-shift on voice identification. Thus, on the contrary to the above, we demonstrate that it is more relevant for voice identification to categorise methods into formant-preserving and non-formant-preserving. WSOLA [18] is the classical example of the latter, while the former is typically exemplified by pitch-synchronous overlap-add (PSOLA) with implementations both in the time domain, TD-PSOLA and the frequency domain, FD-PSOLA [19].

III. Effect of pitch-shift on voice identification

Given the link between long-term formant distributions and speaker identity, we expect that a pitch-shift operation that alters formant frequencies is likely to have a greater impact on speaker identification compared to one that preserves the formants. The x-vector approach to speaker embeddings, based on a deep neural network, represents the state-of-the-art in speaker recognition [1], [2]. The recently introduced ECAPA-TDNN [10] architecture is one-such example that exhibits high speaker recognition performance on established challenge datasets. Below we study the effect of pitch-shift on speaker identification using ECAPA-TDNN x-vectors.

We applied WSOLA, as an example of a non-formant-preserving method of pitch-shift, and TD-PSOLA, as an example of a formant-preserving form of pitch-shift, to a subset of TIMIT utterances [20]. A gender balanced set of 136 male speakers and 136 female speakers was randomly selected from the training partition of TIMIT. With ten utterances available per speaker we randomly-selected three utterances for training (a single combined embedding feature was used to model each speaker, see below) and the remaining seven for prediction. In this way, we can evaluate voice identification accuracy with respect to 517,888 match estimates ($1,904 = 7 \times 272$ positive matches and $51,5984 = (271 \times 7) \times 272$ negative matches).

To assess WSOLA and TD-PSOLA in a consistent manner we define the pitch-shift ratio β. An increase in pitch is denoted by $\beta > 1$ and a decrease is denoted by $\beta < 1$. A value of $\beta = 1$ denotes no pitch-shift. In the case of TD-PSOLA this denotes the multiplicative parameter that is applied to the estimated pitch. That is, for a speech utterance where $z(n)$ is the estimated pitch vector [1] for frame n we define

$$\widehat{z}(n) = \beta z(n) \tag{1}$$

as the modified pitch vector used to synthesize the pitch-shifted speech via the Praat implementation [22] of TD-PSOLA. [2] In the case of WSOLA, we used the implementation available in SoX [24] where the pitch-shift parameter is related to the shift in 100ths of semitones or ‘cents’. This value can be obtained from the pitch-shift ratio as

$$s = 1200 \log_2(\beta) \tag{2}$$

All pitch-shift operations were performed on audio sampled at 16 kHz, matching the sampling rate of the ECAPA-TDNN training data. As we shall explain in the next section, it is also required to establish the impact of pitch-shift at 8 kHz. As such, a resampling operation was applied to the clean and pitch-shifted audio to obtain a dataset sampled at 8 kHz. We extracted x-vectors for all speech utterances, training and prediction. As three x-vectors were made available for training, a single x-vector per speaker model was obtained by component-wise averaging, i.e.

$$x_e = \frac{\sum_i x_i}{||\sum_i x_i||_F} \tag{3}$$

where x_i denotes the embedding extracted for training utterance i and $||\cdot||_F$ the Frobenius norm.

To simulate the effect of a disguised attacker, we applied different degrees of pitch-shift to the prediction utterances using WSOLA and TD-PSOLA and extracted the corresponding x-vectors. Distances were calculated as the dot product between the normalised model and prediction x-vectors, i.e. the cosine distance.

We evaluated the impact of pitch-shift on voice identification via the equal error rate (EER) between the positive and negative distances. Table. I shows the results for the WSOLA and TD-PSOLA methods. As a baseline, the EER for the clean 16 kHz dataset was 0.2% and 0.9% for the 8 kHz dataset. In the case of WSOLA we observe that the more extreme pitch-shift ratios have a large impact on the EER, 37.8% ($\beta = 0.7$, 8 kHz) and 29.5% ($\beta = 1.3$, 8 kHz). However we note, as expected, that the impact to the EER is not as significant for TD-PSOLA, 3.9% ($\beta = 0.7$, 8 kHz) and 3.1% ($\beta = 1.3$, 8 kHz). We can therefore conclude that non-formant-preserving operations such as WSOLA pose a greater threat to speaker identification.

[1] In the case of TD-PSOLA we used [21] to estimate the pitch.

[2] The parselmouth Python wrapper [23] was used to implement Praat.

TABLE I
Pitch-shift and voice identification. Equal error rates (%) for different pitch-shift methods WSOLA (non-formant-preserving) and TD-PSOLA (formant-preserving), pitch-shift ratios (β) and sampling frequencies (8 kHz, 16 kHz). Note the row where $\beta = 1.0$ denotes the baseline equal error rate with no pitch-shift.

β	WSOLA 16 kHz	WSOLA 8 kHz	TD-PSOLA 16 kHz	TD-PSOLA 8 kHz
0.7	37.8	37.8	0.9	3.9
0.8	22.1	22.9	0.6	2.2
0.9	2.5	4.9	0.5	1.7
1.0	0.2	0.9	0.2	0.9
1.1	2.9	5.7	0.5	1.1
1.2	15.5	18.7	0.6	2.5
1.3	28.4	29.5	0.9	3.1

TABLE II
Pitch-shift and voice identification with speaker model augmentation. Equal error rates (%) for WSOLA evasion attacks derived from audio sampled at 8 kHz. The column "none" denotes no augmentation (matches results shown in Table I). Other columns show the impact to EER by augmenting the model embedding, as described in Eqn. (4), using different β sets (B_0, B_1, B_2).

	speaker model augmentation			
β	none	B_0	B_1	B_2
0.7	37.8	36.1	30.6	21.0
0.8	22.9	14.4	7.2	7.1
0.9	4.9	1.7	1.9	3.2
1.0	0.9	1.0	1.4	2.6
1.1	5.7	2.4	2.9	4.5
1.2	18.7	10.7	5.9	5.6
1.3	29.5	24.1	15.4	10.9

In the event a malicious actor is using WSOLA for evasion, the results in Table. I highlight an important concern. Even when small pitch-shifts are employed the EER is still high (5.7% for $\beta = 1.1$ and 4.9% for $\beta = 0.9$ in 8 kHz scenario). Exploring the receiver operating characteristic (ROC) curves for the same set of results – not shown here due to space constraints – the detection rate of small pitch-shifts is low even at a low false positive rate (FPR). For instance it was found that the true positive rate (TPR) is 20% at 0.1% FPR when $\beta = 1.1$. This illustrates how voice identification systems are not equipped to cater for WSOLA-type attacks.

IV. Prevention Strategies

As we have established in the previous section, in the context of protecting voice identification systems from threats, catering for non-formant-preserving pitch-shifts such as WSOLA is paramount as these have the greatest impact on speaker embeddings. We investigate two strategies that can operate in parallel: (i) augmentation of the speaker model embeddings with pitch-shift, and (ii) a pitch-shift-detection model which exploits the discriminative nature of the embeddings themselves.

A. Speaker model augmentation

Data augmentation can better equip systems to cater for unseen data. We thus revisit how the speaker model embedding is calculated in Section III and replace as

$$\widehat{x}_e = \frac{\sum_i x_i + \sum_i \sum_j^B \widehat{x}_{i,\beta_j}}{||\sum_i x_i + \sum_i \sum_j^B \widehat{x}_{i,\beta_j}||_F} \tag{4}$$

where $\widehat{x}_{j,\beta_j}$ denotes embeddings obtained from WSOLA pitch-shift operations of the utterances used for training. We experimented with using different sets of values for augmentation: $B_0 = \{0.9, 1.1\}$; $B_1 = \{0.8, 0.9, 1.1, 1.2\}$ and $B_2 = \{0.7, 0.8, 0.9, 1.1, 1.2, 1.3\}$. In Table II we show how the voice identification EER is impacted by the proposed augmentation strategy in the event of a WSOLA evasion attack. Note we focus specifically on the results corresponding to a 8 kHz sampling frequency. We observe that as we include small changes to pitch in the training data, i.e. augmentation set B_0, it improves the EER for small pitch-changes ($\beta = 0.9$, $\beta = 1.1$). Furthermore this improvement comes at little cost to the EER when the attacker is not using pitch-shift, observe that the EER for $\beta = 1$ increases only from 0.9% to 1.0%. Examining receiver operating characteristic (ROC) curves – not shown due to space constraints – we found B_0 augmentation improves the TPR at 0.1% FPR from 20.4% to 72.1%. This further emphases the gains of B_0 augmentation to mitigate against small pitch-changes while retaining performance levels for instances with no pitch-shift.

On the other hand, while more aggressive augmentation strategies (B_1, B_2) enable reductions in EER for large pitch-shifts (i.e. the EER is reduced from 29.5% to 10.9% for $\beta = 1.3$ using B_2), they come at the cost of increased EER when no pitch-shift is used. In other words, model augmentation via the B_0 strategy helps to address smaller pitch-shifts but more aggressive ones detract from the overall performance of the system. We therefore require a separate strategy for medium and large pitch-shifts.

B. Pitch-shift detection using speaker embeddings

While the sensitivity of speech embeddings to WSOLA-type pitch-shifts is on the one hand a disadvantage, it presents an opportunity to exploit the speaker embeddings themselves as a discriminative feature. We propose a pitch-shift detection model using ECAPA-TDNN x-vectors.

We applied pitch-shift to utterances sampled at 16 kHz, and then resampled all the data to 8 kHz. The motivation for the downsampling operation is to provide reliable results with respect to negative pitch-shifts ($\beta < 1$). As part of negative-pitch WSOLA operations downsampling is applied that leaves an absence of spectral activity in the higher frequencies of the output. This is an artefact that could be easily exploited by a model trained on x-vectors or any features that characterise higher frequencies if using the 16 kHz data. However, due to noise or downsampling that would likely occur post-WSOLA in real-world conditions, this is not an artefact that could be reliably detected and therefore should not be modelled. By downsampling the pitch-shifted and clean speech data we remove this artefact to prevent its usage in training.

For training we employed the training partition of the TIMIT database with 9,240 randomly-selected pitch-shift ratios based on a uniform distribution from the ranges $U(\beta = 0.6, \beta = 0.9)$ and $U(\beta = 1.1, \beta = 1.4)$.[3] The model employed is a low-complexity one comprising three densely-connected layers (128, 32 and 1 units respectively), where the input is the ECAPA-TDNN x-vector and the output is a scalar: 1 if the input is pitch-shifted else 0. Model optimization was performed with respect to the mean square error with mini-batch gradient descent using the RMSProp optimizer.

For comparison, we consider models with similar architectures trained on related feature sets. Our motivation for using x-vectors as a discriminative feature stems from the idea that a voice embedding is a data-driven representation of the speech production model for a given speaker. Previous work on pitch-shift detection [15] has sought to use more classical speech-production features derived from the source-filter model [16]. The so-called speech anomaly detection (SAnD) utilises pitch as well as formants and residual signals derived from LPC analysis. It was shown that such an approach matches the detection performance of human listeners. While in [15] the authors investigate and compare one- and two-class solutions trained using a SVM, we here wish to focus on the considered features only. For the same training dataset as above we extracted the mean, standard deviation, median and median absolute deviation of the pitch, the first two formants, and the kurtosis of the LPC-residual over time giving a 16×1 feature vector per speaker. [4] Accounting for the difference in the input dimension, we trained a SAnD-model using the same architecture as the proposed one.

MFCCs are a concise representation of acoustic features widely used in a range of speech applications [26]. We obtained 20 MFCCs for overlapping frames[5] for the same training dataset used by both the proposed x-vector approach and SAnD. We calculated the mean and standard deviation of both the coefficients and the delta coefficients, i.e. the first derivative, along time. Discarding the low-frequency coefficient, this yields a 76×1 feature vector per speaker. Again accounting for the difference in the input dimension, we trained an MFCC-model using the same proposed architecture.

In Table III we compare the EERs obtained using SAnD features, MFCC features and x-vectors for different degrees of pitch-shift. Testing was conducted on the 1680 utterances from the test partition of the TIMIT database. Clearly the EER is lowest using x-vectors for each considered pitch-shift ratio. Another observation is that for each of the considered feature sets, the detection performance is lowest for pitch-shift ratios close to one ($\beta = 0.9$, $\beta = 1.1$); this result is consistent with previously-reported findings [15]. For moderate ($\beta = 0.8$, $\beta = 1.2$) and more extreme pitch-shifts ($\beta = 0.7$, $\beta = 1.3$) the detection rates are much higher with x-vector EERs $\leq 2.1\%$.

TABLE III
Pitch-shift detection. EQUAL ERROR RATES (%) IN THE DETECTION OF WSOLA PITCH-SHIFTS USING DIFFERENT FEATURES FOR SIX PITCH-SHIFT RATIOS β APPLIED TO THE TEST PARTITION OF THE TIMIT DATASET (RESAMPLED TO 8 kHz AFTER PITCH-SHIFT OPERATIONS).

β	SAnD	MFCCs	x-vectors
0.7	9.7	7.5	**0.9**
0.8	19.7	9.6	**1.2**
0.9	36.3	23.5	**11.3**
1.1	38.0	29.0	**13.7**
1.2	23.7	11.8	**2.1**
1.3	15.6	9.2	**1.0**

We next tested the x-vector approach on a larger dataset of 104,014 utterances from the 360 partition of the LibriSpeech corpus [27] to evaluate performance for a wider range of operating thresholds. Embeddings were obtained from WSOLA-shifted utterances resampled to 8 kHz from 16 kHz as described previously. We experimented with adjusting the range of the training data from $U(\beta = 0.6, \beta = 0.9)$ and $U(\beta = 1.1, \beta = 1.4)$ to $U(\beta = 0.6, \beta = 0.8)$ and $U(\beta = 1.2, \beta = 1.4)$ to explore impact on detection performance. ROC curves – not shown due to space constraints – reveal that removing small pitch-changes from the training data, i.e. $U(\beta = 0.6, \beta = 0.8)$ and $U(\beta = 1.2, \beta = 1.4)$, enables better detection rates for operating points, e.g. FPRs $< 1\%$.

C. Combined system

We evaluated the gains of a malicious attacker detection system using 104,014 utterances from the 360 partition of the LibriSpeech corpus [27]. Embeddings were obtained from WSOLA-shifted utterances resampled to 8 kHz from 16 kHz as described previously. We randomly selected six speakers (three female) as malicious actors. For each malicious actor, we assumed 10 of the original utterances were available for training and the remaining utterances were unseen attacks for testing (612 total across all speakers). Large ($\beta = 1.3$) and small ($\beta = 1.1$) pitch increases were applied to the attack utterances such that the total number of attack utterances was 1,224. All other 103,342 utterances from the dataset were used as the negative set.

Standard (no augmentation) and augmented speaker models were trained on each of the six malicious actors;[6] the voice identification score was taken as the maximum score from the models. The combined system score is a weighted average of the voice identification score and the pitch-detection score.[7] In Fig. 2 we show the ROC curves using only the standard voice identification system, and combined systems (voice identification and pitch-shift detection) with and without augmentation. It is clear how the standard voice identification system fails detect the vast majority of attacks at operating points corresponding to FPRs below 1%. As expected, the combined systems fare better. In particular the advantage of

[3] Note we balanced the training data with respect to gender.

[4] With respect to the SAnD features, a frame size of 20 ms was used for the LPC analysis yielding formant estimates and the LPC residual signal. A frame size of 40 ms was used for pitch estimation using YIN [25].

[5] MFCCs were calculated using a frame size of 40 ms with 20 ms overlap.

[6] The speaker models was augmented with the B_0 set.

[7] The pitch-detection model was trained on the TIMIT corpus, as described in Section IV-B, with $U(\beta = 0.6, \beta = 0.8)$ and $U(\beta = 1.2, \beta = 1.4)$.

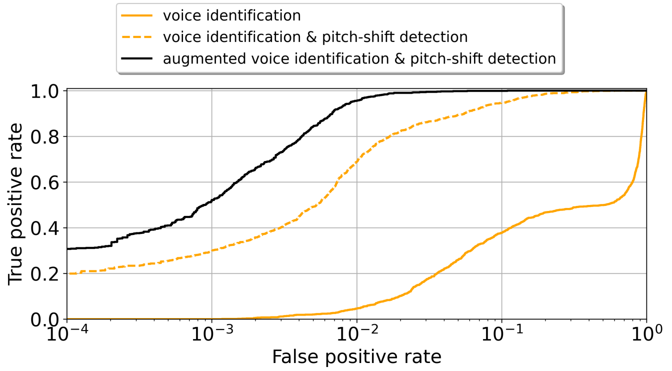

Fig. 2. **Combined System.** ROC curves for large ($\beta = 1.3$) and small ($\beta = 1.1$) pitch-shift attacks using a standard voice identification system, and dual-systems comprising (i) voice identification and pitch-shift detection and (ii) augmented voice identification and pitch-shift detection.

augmenting the speaker models is clear; it facilitates a 20% absolute increase in detection at an FPR of 0.1% within the combined framework.

V. Conclusions

In the evaluation of the threat posed by pitch-shift to voice recognition, specifically voice identification, we have established that the main concern encompasses operations that do not preserve formant information. We have shown that even small changes in pitch ($\beta = 1.1$, $\beta = 0.9$) using such methods facilitate evasion. A dual-solution strategy is proposed. On the one hand, we show how a pitch-detection model, exploiting the sensitivity of speech embeddings to such attacks, can reliably detect medium and large changes in pitch ($\beta \leq 0.8$, $\beta \geq 1.2$). On the other hand, for smaller changes in pitch, augmentation of the training data of the voice models is a more appropriate solution.

References

[1] E. Khoury and M. Garland, "End-to-end speaker recognition using deep neural network," US Patent 9,824,692, 2017.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 5329–5333.

[3] J. H. Künzel, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Effect of voice disguise on the performance of a forensic automatic speaker recognition system," in *Odyssey 2004 - The Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004.

[4] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection: Review and perspectives," in *Progress in Nonlinear Speech Processing*, Y. Stylianou, M. Faundez-Zanuy, and A. Esposito, Eds. USA, NY: New York:Springer-Verlag, 2017, pp. 101–117.

[5] M. Farrús, "Voice disguise in automatic speaker recognition," *ACM Comput. Surv.*, vol. 51, no. 4, jul 2018. [Online]. Available: https://doi.org/10.1145/3195832

[6] Y. Wang and Z. Su, "Detection of voice transformation spoofing based on dense convolutional network," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 2587–2591.

[7] P. Perrot and G. Chollet, "The question of disguised voice," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3878–3878, May 2008. [Online]. Available: https://doi.org/10.1121/1.2935782

[8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, Graz, Austria, Sep 2019.

[9] X. Liu, X. Wang, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2507 – 2522, 2023.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[11] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 824–837, 2021.

[12] H. Wu, Y. Wang, and J. Huang, "Identification of electronic disguised voices," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 489–500, 2014.

[13] W. Cao, H. Wang, H. Zhao, Q. Qian, and S. M. Abdullahi, "Identification of electronic disguised voices in the noisy environment," in *Digital Forensics and Watermarking*, Y. Q. Shi, H. J. Kim, F. Perez-Gonzalez, and F. Liu, Eds. Cham: Springer International Publishing, 2017, pp. 75–87.

[14] Y. Ye, L. Lao, D. Yan, and R. Wang, "Identification of weakly pitch-shifted voice based on convolutional neural network," *Intl. J. Digital Multimedia Broadcasting*, 2020. [Online]. Available: https://doi.org/10.1155/2020/8927031

[15] D. Looney and N. D. Gaubitch, "On the detection of pitch-shifted voice: Machines and human listeners," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada, Jun. 2021, pp. 5804–5808.

[16] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[17] X. Li, Z. Zheng, C. Yan, C. Li, X. Ji, and W. Xu, "Toward pitch-insensitive speaker verification via soundfield," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 1175–1189, 2024.

[18] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, 1993, pp. 554–557 vol.2.

[19] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, 1990.

[20] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.

[21] M. Morrison, C. Hsieh, N. Pruyne, and B. Pardo, "Cross-domain neural pitch and periodicity estimation," in *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 2023.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/, 2021.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] "SoX sound eXchange," http://sox.sourceforge.net/Main/HomePage, accessed: 2024-03-10.

[25] A. de Cheveigné and H. Kawahara, "YIN, A fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, pp. 1917–1930, 2002.

[26] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, aug 1980.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5206–5210.