Applied Acoustics 183 (2021) 108283

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Towards protecting cyber-physical and IoT systems from single- and multi-order voice spoofing attacks

Ali Javed [a,b], Khalid Mahmood Malik [a,*], Aun Irtaza [a], Hafiz Malik [c]

[a] Department of Computer Science and Engineering, Oakland University, Rochester 48309-4479, MI, USA
[b] Department of Computer Science, University of Engineering and Technology-Taxila, 47050, Pakistan
[c] Department of Electrical and Computer Engineering, University of Michigan-Dearborn, 4901 Evergreen Road, Dearborn 48128, MI, USA

ARTICLE INFO

Article history:
Received 17 December 2020
Received in revised form 13 April 2021
Accepted 28 June 2021
Available online 17 July 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Cloned-replay attack
Cyber-physical systems
Voice replay attack
Voice cloning attack
Voice spoofing countermeasure

ABSTRACT

Voice-controlled systems (VCSs), a new class of cyber-physical systems (CPS), and Internet of Things (IoT) devices are increasingly employing smart speakers such as Google Home and Amazon Alexa, and other voice assistants to enable management of various remote operations at home and offices. However, these smart speakers and hence VCSs are susceptible to various voice spoofing attacks i.e. replay, cloning, etc., in a non-network environment as well as in a multi-hop network setup. These diverse spoofing threats on VCSs require an urgent need to develop a robust spoofing countermeasure for VCSs capable of detecting a variety of voice spoofing attacks. This paper presents a spoofing countermeasure that uses novel acoustic ternary patterns (ATP) with Gammatone cepstral coefficients (GTCC) features to counter the voice spoofing attacks on VCSs in single- and multi-hop network environments. Our experimental analysis demonstrates that the proposed ATP features when combined with GTCC can effectively detect the distortions in replayed samples, unnatural prosody present in the cloned samples, and both distortions and unnatural patterns of stress and intonation in cloned-replay samples. The proposed ATP-GTCC features are used to train the SVM for development of a spoofing countermeasure to cater all possible forgeries. Experimental results based on highly diversified ASVspoof 2019 and VSDC datasets signify the effectiveness of the proposed countermeasure for reliable detection of 1st- and 2nd-order replay, cloning, and cloned-replay attacks.

1. Introduction

The feature of smart speakers (e.g. Amazon Echo, Google Home) to control various home/office appliances and actuators etc., is making them an essential component of Internet of Things (IoT) and Cyber-physical systems (CPS). Although smart speakers and other voice assistants, acting as interface of voice-controlled systems (VCSs), have transformed the IoT and CPS domains, but also resulted in the generation of new potential threats. For example, impostors may retrieve sensitive data from healthcare and/or financial applications through executing the voice spoofing attacks on VCSs to commit financial frauds [1], or to gain unwanted remote access of smart homes and offices [2]. Additionally, Covid-19 crisis is expected to accelerate the use of voice as an authentication mechanism for many businesses and service industry, as other authentication mechanism (e.g. use of keypad, finger scan, etc.) could cause transmission of infection.

Existing voice spoofing attacks, also known as voice presentation attacks [3] i.e. replays, cloning, etc., can easily be used to spoof the VCSs. Voice replays are generated by playing the recorded voice of the actual speaker with the intention of deceiving the VCS. Voice cloning, which refers to creating a synthetic voice of the target speaker, can easily be generated due to the advancement of sophisticated deep learning algorithms.

Various spoofing countermeasures (CMs) have been proposed to detect either replay, cloning, or both attacks. In [4], constant Q-transform cepstral coefficients (CQCC) were used, whereas, in [5], constant-Q variance-based octave coefficients and constant-Q mean-based octave coefficients were used with Gaussian mixture model (GMM) for replay attack detection. Some spoofing CMs [6,7] have examined high-frequency bands of the audio using different features including the amplitude and frequency-based modulation [6] and various cepstral coefficients features [7] with the GMM for replay spoofing detection. These methods [6,7] provide

[*] Corresponding author.
E-mail address: mahmood@oakland.edu (K.M. Malik).

https://doi.org/10.1016/j.apacoust.2021.108283
0003-682X/© 2021 Elsevier Ltd. All rights reserved.

better detection performance over the baseline model [4], however with increased features computation cost. Existing replay spoofing CMs [8,9] have also used deep learning models. In [8], MFCC and CQCC features were used with hybrid classification model consisting of GMM, deep neural network and ResNet to classify between the genuine/bonafide and replay signal. In [9], deep generative variational auto-encoder framework was employed for replay detection.

Existing techniques [10–16] have used different magnitude- and phase-based features for voice cloning/synthesis detection. In [10,11], relative phase shift features were used, whereas, cochlear filter cepstral coefficients (CFCC) and CFCC-instantaneous frequency features were employed in [12] with the GMM for voice cloning detection. In [13], inverted constant-Q coefficients, inverted CQCC, inverted C-Q block coefficients, and inverted C-Q linear block coefficients were employed for speech synthesis detection. In [14], a non-learning technique based on higher-order spectral features was employed for voice cloning detection. Although, this method [14] achieves better performance, but has a high features computation cost. In [15], an end-to-end ensemble model was proposed to detect the replay and cloning attacks. Light convolutional neural network based on angular margin-based softmax activation function was proposed in [16] for voice replay and cloning attacks detection.

In our prior work [4], we proved that the latest VCSs are susceptible to replay attacks and can be exploited to cause severe damages in cyber-physical systems e.g. home and office automation control. Further, we also validated in [4] that Google Home and Amazon Alexa devices are vulnerable to multi-order replay attacks even in multi-hop/chained scenarios. Fig. 1(b) shows chained VCS scenario that generate multi-order replays when input is human speaker or its cloned voice. It occurs when one VCS replays the genuine or synthetic voice to next connected one in the chain. For example, a hacker uses his cellphone to replay the recorded voice command of human speaker e.g.“Alexa, turn off the heat” (1st-order replay) on the baby monitor (VCS-2) that is accessed by invading the wireless LAN using tools i.e.Aircrack. Next, the voice command is replayed (2nd-order replay) to the VCS of targeted person’s home (VCS-3) to turn off the heat. Unlike traditional applications which consider spoof detection as a binary problem, we consider this as a multiclass problem for chained VCSs, because, it is possible for a certain VCS, which itself has robust binary spoof detection mechanism, to receive cloned or playback voice from other VCSs that are either compromised or prone to voice spoofing attacks due to a weak or absent spoof detection approach. Thus, the received audio will be considered genuine and the spoofing detector will ultimately fail for all the chained devices. Moreover, this work introduces *a new voice spoofing threat i.e. cloned-replay* that can also be generated in multi-hop scenarios. We generated the 1st- and 2nd-order cloned-replay recordings using the cloning samples of ASVspoof 2019 LA dataset. Similarly, our analysis shows that these VCSs are also vulnerable to voice cloning attacks. As shown in Fig. 1(a), a cloned voice command is directly played on the VCS to generate a single-order cloning attack to control the lighting system.

Literature shows that countermeasures trained with one class of spoofing attacks fail to generalize well for other classes of spoofing attacks [25,26]. For example, systems trained with speech cloning show poor performance for replay detection [27]. Additionally, the existing methods are not designed for detection of multi-order attacks either. Therefore, there is need to develop model which can capture microphone induced distortions in the replay/playback samples, differentiate high-order distortions when same voice is played back in the chain, and the natural pauses of human model of speech production are missing from the synthetic voice generated by deep learning based speech cloning algorithms. To meet these requirements of VCSs which are connected via single- or multi-hop IoT and CPS networks, we propose a robust spoofing countermeasure that can reliably detect the various single-order and multi-order (i.e. chained) voice spoofing attacks (i.e. replay, cloned, and cloned-replay) using the proposed acoustic-ternary patterns (ATP). The human speaker’s speech has dynamic characteristics due to speaker induced variations, whereas, the cloned voice generated by various state-of-art deep learning-based speech synthesis algorithms [24] contains unnatural prosody such as emphasis on the wrong syllables or words, absence of natural pauses, lack of unvoiced consonants, unnatural pitch, and small percentage of mispronunciation. This introduction of deviation in patterns of rhythm and sound in synthesized speech calls for capturing time-domain specific aspects in generated speech.

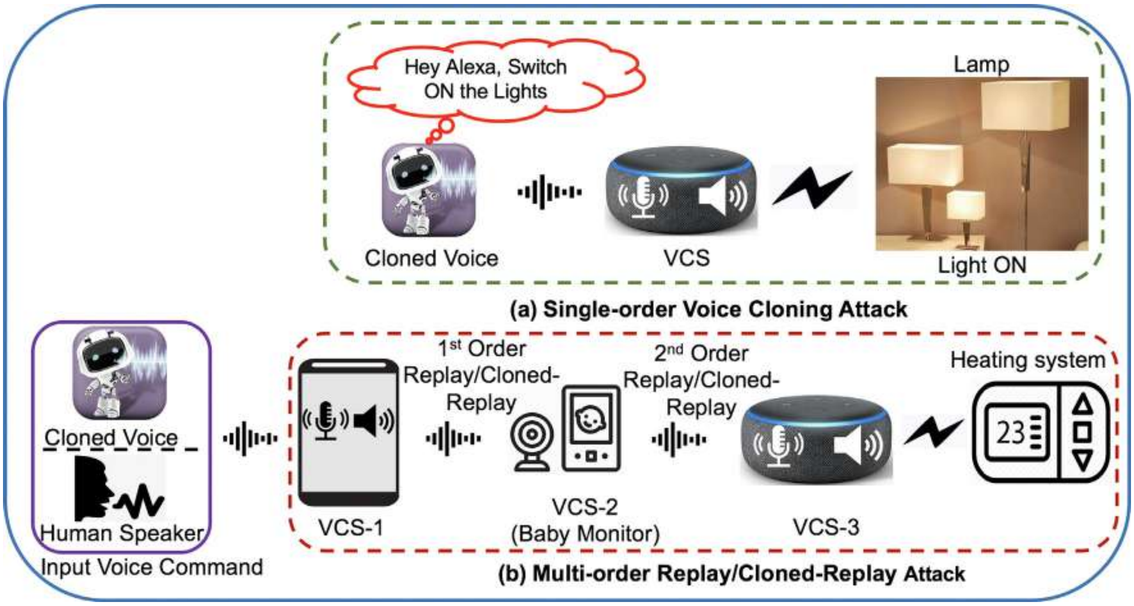
Hey Alexa, Switch ON the Lights
Cloned Voice
VCS
Lamp
Light ON
(a) Single-order Voice Cloning Attack
Cloned Voice
Human Speaker
Input Voice Command
VCS-1
1st Order Replay/Cloned-Replay
VCS-2 (Baby Monitor)
2nd Order Replay/Cloned-Replay
VCS-3
Heating system
23
(b) Multi-order Replay/Cloned-Replay Attack

Fig. 1. Voice spoofing attacks scenarios.

Therefore, we propose ATP features to better capture these dynamic attributes of speech variations between genuine and synthesized speech. Moreover, replay samples contain the microphone induced distortions, and cloned samples contain robotic 'whine' which can be effectively captured by both the ATP and the gammatone cepstral coefficients (GTCC) due to their robustness against the environmental noise. Therefore, we fused the ATP with the GTCC features to develop a unified countermeasure to detect multi-order replay, cloning, and cloned-replay forgeries. Moreover, since these distortions increase in higher-order replays therefore our proposed features can provide even better detection performance in such scenarios. The main contributions of this paper are as follows:

- We propose a novel acoustic ternary pattern features for audio representation to differentiate spoofed samples from the bonafide.
- We present the groundwork for a spoofing countermeasure capable of detecting the multi-order replay, cloning, and cloned-replay attacks using the proposed ATP-GTCC features.
- We present a new voice spoofing attack (cloned-replay) that can be generated by replaying the cloned audios.
- We performed rigorous experimentation against existing state-of-the-art voice spoofing countermeasures to indicate the effectiveness of the proposed method.

2. Proposed method

This section provides a discussion on the proposed spoofing countermeasure to detect multiple voice spoofing attacks. The architecture of our spoofing countermeasure is shown in Fig. 2.

2.1. Features extraction

To better capture the traits of multiple voice spoofing attacks, we propose a novel ATP feature descriptor and fused it with GTCC to represent the audio. The details of the features extraction process are presented in the subsequent sections.

2.1.1. Acoustic ternary patterns (ATP)

This paper presents a novel features representation scheme, acoustic ternary patterns for audio signals representation. For a given audio signal $X = Y^{(i)}[m]_{i=1}^{i=M_f}$, we partition it into M_f non-overlapping frames $Y^{(i)}[m]$ having frame length of L. Let $Y^{(i)}[j]$ represents the central sample having N^k neighboring samples, and k represent the neighbor index against the central sample. To compute the acoustic ternary code for each frame, we calculate the magnitude difference of the signal between $Y^{(i)}[j]$ and neighboring samples (N^k) by applying the threshold t. To compute the value of t, we employ a linear search scheme where we initialize the t to zero and optimize it to search the convergence point in the range of 0 to 1. For our experiments, t = 0.00015 provides the most accurate results. Next, we quantize the signal values in the range of $\pm t$ around $Y^{(i)}[j]$ to zero, whereas the values above $Y^{(i)}[j] \pm t$ are quantized to 1 and below $Y^{(i)}[j] \pm t$ to 1. Thus, we get a three valued ATP code as:

$$F(N^k, Y^{(i)}[j], t) = \begin{cases} -1, & N^k - (Y^{(i)}[j] - t) \leqslant 0, \\ 0, & (Y^{(i)}[j] + t) < N^k < (Y^{(i)}[j] - t) \\ +1, & N^k - (Y^{(i)}[j] + t \geqslant 0) \end{cases} \quad (1)$$

where $F(N^k, Y^{(i)}[j], t)$ represents the speech signal using a three valued ternary code/pattern. Later, we split the patterns (F) into upper (F^{up}) and lower (F^{low}) patterns. We retain all values quantized to +1 in (F^{up}) and replace all remaining values with zeros as follows:

$$F^{up}(N^k, Y^{(i)}[j], t) = \begin{cases} 1, & F(N^k, Y^{(i)}[j], t) = +1 \\ 0, & Otherwise \end{cases} \quad (2)$$

For lower patterns, we retain all values quantized to -1 in F^{low} and replace the other values with zeros as:

$$F^{up}(N^k, Y^{(i)}[j], t) = \begin{cases} 1, & F(N^k, Y^{(i)}[j], t) = -1 \\ 0, & Otherwise \end{cases} \quad (3)$$

Thus, we can represent the ternary patterns as:

$$TP(Y^{(i)}[j]) = \begin{cases} \sum_{r=-1}^{r=1} \sum_{s=-1}^{s=1} F^{up}[(Y^{(i)}[r,s] - Y^{(i)}[j])\| \\ \ldots \sum_{r=-1}^{r=1} \sum_{s=-1}^{s=1} F^{low}[(Y^{(i)}[r,s] - Y^{(i)}[j]) \end{cases} \quad (4)$$

Next, we obtain the uniform patterns for acoustic features representation as uniform patterns represent primitive information and hold maximum attributes of the signal over non-uniform patterns [17]. More specifically, we computed the upper uniform and lower uniform ternary patterns from the $F^{up}(.)$ and $F^{low}(.)$, and transformed these binary representations of patterns into decimal form as follows:

$$U(TP^{up}) = \sum_{k=0}^{k=7} F^{up}(N^k, Y^{(i)}[j], t) \times 2^k \quad (5)$$

$$U(TP^{low}) = \sum_{k=0}^{k=7} F^{low}(N^k, Y^{(i)}[j], t) \times 2^k \quad (6)$$

where the U value of the ternary patterns represents the number of bitwise transitions (0/1 changes) in the pattern, and those having minimal discontinuities are denoted as uniform, i.e. 00000000 and 01000000 patterns have U values of 0 and 2, respectively. During histogram encoding, we can considerably reduce the histogram bins by assigning all non-uniform patterns to a single bin without losing significant information. For this, we computed the histogram of TP^{up} and TP^{low}, and assigned one histogram bin for each uniform pattern, placing all non-uniform patterns in a single bin. Histograms are calculated as:

$$H_{T_p}^{up}(TP^{up}, b) = \sum_{q=1}^{Q} \delta(TP_q^{up}, b) \quad (7)$$

$$H_{T_p}^{low}(TP^{low}, b) = \sum_{q=1}^{Q} \delta(TP_q^{low}, b) \quad (8)$$

where b represents the histogram bins corresponding to the uniform TP codes, and $\delta(.)$ is the Kronecker delta function. After performing extensive experiments, our analysis showed that the first ten upper- and lower-uniform patterns each were enough to capture all traits available in the genuine and spoof samples, as we didn't achieve any performance improvement when used more patterns. Thus, we concatenated the 10-D $H_{T_p}^{up}$ and 10-D $H_{T_p}^{low}$ to create a 20-D ATP features descriptor as:

$$ATP = [H_{T_p}^{up} || H_{T_p}^{low}] \quad (9)$$

ATP Features Analysis. As the voice interfaces of VCSs and IoT devices are susceptible to replay, cloning, and cloned-replay attacks, therefore, an effective spoofing detector/countermeasure should take into account these facts during feature extraction: (1) The microphone adds a layer of non-linearity due to

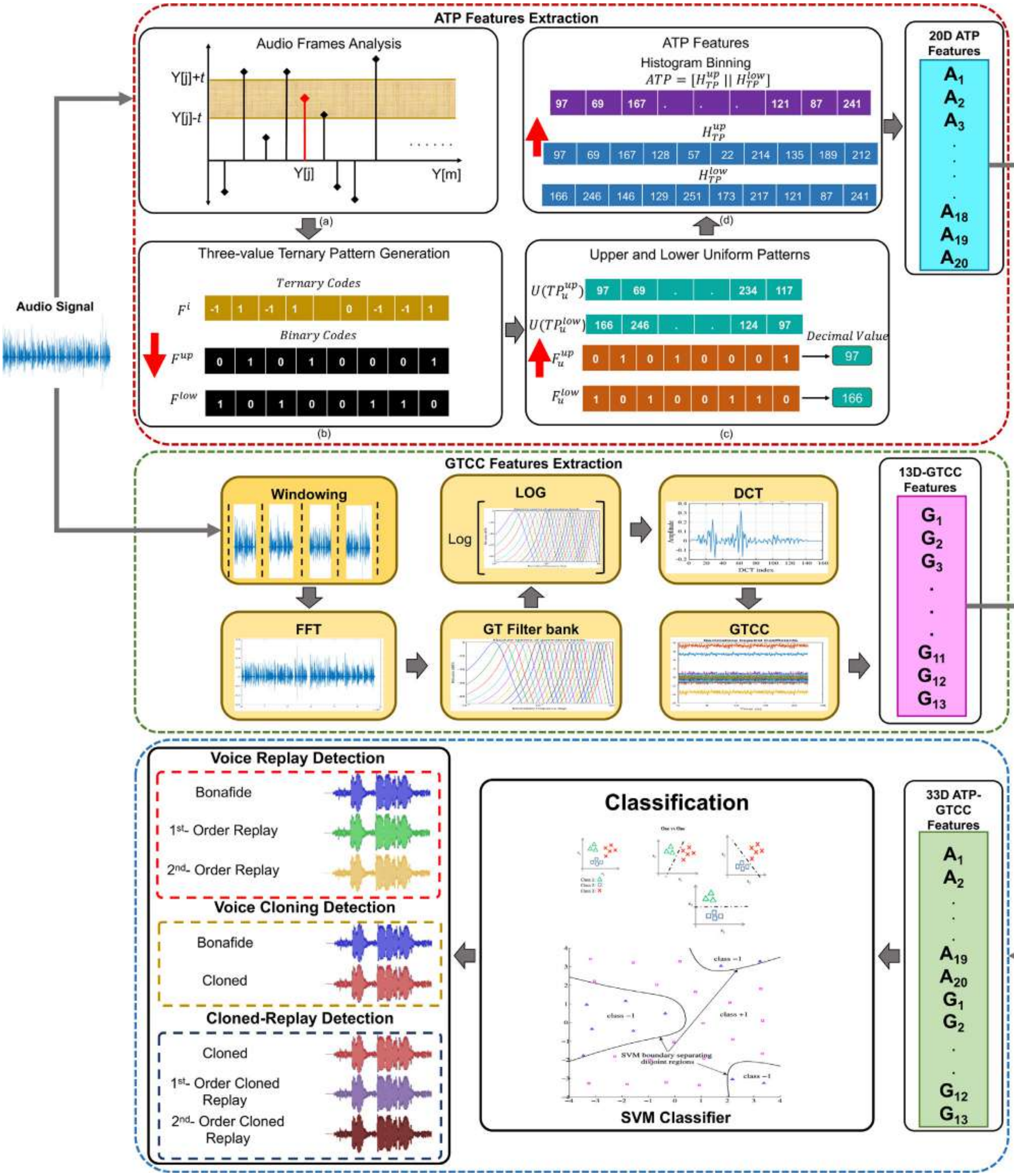

Fig. 2. Architecture of proposed method.

inter-modulation distortions, which induce the detectable patterns [19], thus must be exploited to develop an effective countermeasure, (2) The subsequent recordings of the same recording thus bring higher-order non-linearities in the audio to make it more detectable, (3) Audio synthesis algorithms also introduce certain artifacts. Thus, all these observations can be utilized to develop a noise resistant and robust countermeasure for real-time applications. As the proposed ATP method analyzes the patterns of the audio, thus effectively captures these artifacts to differentiate between the bonafide and spoofed audios. To suppress the additive noise, upper and lower threshold values are also considered with the central sample $Y^{(i)}[j]$ in M_f that also lowers the probability of wrong ternary code generation. Furthermore, less complex ATP features enable fast model retraining; thus, makes our technique efficient for applications involving automatic speaker verification with continuous user enrollment.

2.1.2. Gammatone cepstral coefficients (GTCC)

GTCC features [18] can be used to capture the distortions in the frequency scale of the audio. Since the voice spoofing datasets contain audio samples collected in noisy environments, we need a descriptor capable of effectively capturing the traits of spoof signals under noisy conditions. GTCC is more robust to noise [18] over other spectral features i.e. MFCC, therefore, we used GTCC features along-with ATP for audio representation. To extract the GTCC features, we computed the fast Fourier transform (FFT) of the audio. Next, we computed the energy of each sub-band E_n by applying the gammatone filter bank on the FFT of signal. The logarithm of each energy band E_n is computed, and discrete cosine transform is employed to extract the GTCC features as:

$$GTCC_p = \sqrt{\frac{2}{z}\sum_{z=1}^{Z} \log E_n \cos\left[\frac{\pi z}{Z}\left(p - \frac{1}{2}\right)\right]} \tag{10}$$

where E_n, Z, and P represents the signal energy for nth spectral band, number of gammatone filters and number of GTCC respectively. We obtain a 13-dimensional GTCC vector using a window length of 30 ms and overlap factor of 20 ms. Finally, we fused the 13D GTCC features with the 20D ATP features to create a 33D ATP-GTCC feature vector to effectively capture the attributes of genuine and spoof samples. The implementation of the proposed features can be found at [20].

2.2. Classification

For classification, we employed the multi-class SVM classifier to distinguish between the bonafide, 1st-order, and 2nd-order spoofing samples. More specifically, replay and cloned-replay spoofing are detected for both 1st- and 2nd-order, whereas, for voice cloning we trained a binary SVM classifier to detect the genuine and cloned samples. We selected the SVM in the proposed method due to the following reasons: 1) SVM has a property of convex optimization which helps to achieve optimal results via global minimum rather than the local minimum. 2) SVM is also effective in case of imbalance dataset. Since the ASVspoof 2019 dataset contains more spoofing samples as compared to the bonafide ones (Table 1), therefore, SVM was considered to handle the data imbalance problem for the ASVspoof 2019 dataset. We employed the proposed features to train the SVM for classification. We tuned the SVM using different kernels and achieved best results with higher-order polynomial kernel (cubic). Therefore, we used the SVM tuned with the cubic kernel. We tuned two parameters during the experiments. The penalty parameter a.k.a box constraint was set to 1, and the kernel scale a.k.a gamma was set to 1.4, as we achieved optimal results with these parameter settings.

3. Experimental results and discussion

3.1. Dataset

We evaluated the performance of the proposed spoofing countermeasure on our voice spoofing detection corpus (VSDC) [21] and the ASVspoof 2019 dataset [22]. VSDC consists of two main collections, one contains the 1st- and 2nd-order replays against the genuine samples, whereas the second contains the 1st- and 2nd-order replays of cloned samples. Our VSDC replays collection is more diverse in comparison of the ASVspoof 2019 corpus as VSDC contains both the 1st- and 2nd-order replay samples against the genuine audios. The details of the VSDC can be found at [21]. ASVspoof 2019 dataset comprises of two collections i.e. PA (replay) and LA (cloning); both collections contain the training, development (dev), and evaluation (eval) sets. The training set contains 54000 and 25380 samples; the dev set contains 33534 and 24844 samples; and the eval set contains 153522 and 71933 samples; from the PA and LA collections, respectively.

3.2. Performance evaluation of proposed countermeasure

Performance of our voice spoofing countermeasure is evaluated using the min-tDCF, equal error rate (EER), precision, recall, f1-score, and accuracy. For replay attacks, we evaluated the results on both VSDC and ASVspoof 2019 datasets. Whereas, we used only the ASVspoof 2019 corpus to evaluate the performance of speech synthesis and VSDC for cloned-replay detection. For VSDC, we used 70% samples for training and remaining 30% for testing, whereas, for the ASVspoof 2019 corpus, we used the training and eval sets for model training and testing respectively. The details of dataset partitioning for experiments are shown in Table 1.

3.2.1. Performance evaluation of ATP and GTCC features

We designed a multi-stage experiment to examine the performance of ATP, GTCC, and ATP-GTCC features fusion for voice spoofing detection. First, we used the proposed ATP features to train the SVM on VSDC and ASVspoof 2019 datasets individually and results are shown in Table 2. Next, we repeat the same with GTCC features. Finally, we employed the ATP-GTCC features for voice spoofing detection. From Table 2, we can observe that the ATP provides better detection performance over GTCC, however, the fusion of ATP and GTCC outperforms both the ATP and GTCC features alone. Therefore, we employed the ATP-GTCC features with SVM to detect the replay, cloning, and cloned-replay attacks.

3.2.2. Performance evaluation of the proposed ATP-GTCC features for multiple voice spoofing detection

We designed an experiment to select the best performing kernel for our SVM. For this purpose, we employed the proposed ATP-GTCC features to train the SVM using different kernels on both VSDC and ASVspoof 2019 datasets. The results are shown in Table 3. We can observe from the results that the SVM with higher-order polynomial kernel (cubic) provides better classification performance over other kernels. More specifically, we achieved an EER of 0.6% on VSDC and 1.1% on ASVspoof dataset for replay detection. Whereas, obtained an EER of 0.1% and 0.09% for cloning and cloned-replay detection respectively. This signify the effectiveness of higher-order polynomial kernel for detecting the distortions in 1st- and 2nd-order spoofing samples. Radial basis function (RBF) kernel achieves second best results by a small margin. Whereas, SVM tuned with linear kernel performs the worst. Thus, we claim that SVM tuned with the cubic kernel using our ATP-GTCC features effectively classifies the genuine and spoof samples.

Table 1
Dataset partition for experiments.

Dataset	Training		Testing	
	Sample set	No. of samples	Sample set	No. of samples
ASVspoof-LA	Train set	25,380	Evaluation set	71,933
ASVspoof-PA	Train set	54,000	Evaluation set	1,53,522
VSDC	70%	8397	30%	3603

Table 2
Performance evaluation of ATP, GTCC, and ATP-GTCC features.

Spoofing Type	Dataset	Features	min-tDCF	EER%	Precision%	Recall%	Accuracy%
Replay	VSDC	ATP	0.194	2.9	96.8	97.4	97
		GTCC	0.497	7.5	91.4	93.5	92.4
		ATP-GTCC	0.04	0.6	99.3	99.3	99.4
	ASVspoof	ATP	0.24	3.4	96.4	96.6	96.5
		GTCC	0.561	8.4	91.2	92	91.5
		ATP-GTCC	0.069	1.1	99.25	99.25	99.2
Cloning	ASVspoof	ATP	0.06	0.9	99	99.1	99
		GTCC	0.42	6.1	93.8	94.3	94
		ATP-GTCC	0.015	0.1	99.9	99.9	99.9
Cloned-Replay	VSDC	ATP	0.072	1.2	98.6	99	98.9
		GTCC	0.29	4.1	96	96	96
		ATP-GTCC	0.014	0.09	99.9	99.9	99.9

Table 3
Voice spoofing detection of the proposed method on different SVM kernels.

Spoofing Type	Dataset	SVM Kernel	EER%	Precision%	Recall%	F1-Score%	Accuracy%
Replay	VSDC	Linear	18	82	82	82	82.2
		Quadratic	1.16	98.3	98.3	98.3	98.3
		Cubic	0.6	99.3	99.3	99.3	99.4
		RBF	0.6	99.3	99.3	99.3	99.4
	ASVspoof	Linear	2	93.47	93	93.23	93.1
		Quadratic	1.5	98.5	98.5	98.5	98.8
		Cubic	1.1	99.25	99.25	99.25	99.2
		RBF	1	99	99	99	99.1
Cloning	ASVspoof	Linear	0.5	99.4	99.5	99.45	99.4
		Quadratic	0.3	99.6	99.8	99.7	99.6
		Cubic	0.1	99.9	99.9	99.9	99.9
		RBF	0.15	99.8	99.9	99.85	99.9
Cloned-Replay	VSDC	Linear	0.44	99.6	99.6	99.6	99.6
		Quadratic	0.2	99.8	99.8	99.8	99.7
		Cubic	0.09	99.9	99.9	99.9	99.9
		RBF	0.15	99.8	99.9	99.85	99.8

3.2.3. Performance comparison of proposed features with different features-combinations

To evaluate the effectiveness of the proposed features for spoofing detection, we generated different feature-combinations of ATP and spectral (i.e. ATP-GTCC, ATP-MFCC, and MFCC-GTCC). For classification, we used the SVM tuned with cubic kernel and results are shown in Table 4. From these results, we can conclude that our ATP-GTCC features outperform other features by achieving the lowest min-tDCF and EER. More precisely, we achieved the min-tDCF and EER of 0.04 and 0.6% on VSDC, whereas, 0.069 and 1.1% on ASVspoof 2019 dataset respectively. Similarly, we achieved the lowest min-tDCF and EER of 0.015 and 0.1% for cloning, whereas, 0.014 and 0.09% for cloned-replay attacks. It is important to mention that the fusion involving ATP features achieved better results over spectral features fusion (i.e.MFCC-GTCC). This shows the effectiveness of ATP features for voice spoofing detection.

3.2.4. Performance comparison of proposed features on different classifiers

To measure the effectiveness of SVM over other classifiers for voice spoofing detection, we performed a comparative analysis of SVM against other classifiers such as k-nearest neighbor (KNN), naïve bayes, decision trees, ensemble bagged trees, and BiLSTM deep learning model. For this experiment, we employed the proposed features to train all of these classifiers separately and results are presented in Table 5. We followed the same experimentation protocol as done for other experiments. From the results of this experiment, we observed that SVM performs the best and Naïve Bayes is the worst for all three spoofing categories. More specifically, SVM achieves the lowest min-tDCF and EER of 0.04 and 0.6%, 0.01 and 0.15%, and 0.006 and 0.09%, whereas Naïve Bayes achieves the highest min-tDCF and EER of 0.528 and 27%, 0.098 and 2.8%, and 0.072 and 1.41% for replay, cloning, and

Table 4
Comparative analysis of different features combination for voice spoofing detection.

Dataset	Features	Replay		Cloning		Cloned-Replay	
		min-tDCF	EER%	min-tDCF	EER%	min-tDCF	EER%
ASVspoof	MFCC-GTCC	0.63	9.25	0.21	3.1	0.19	2.8
	ATP-MFCC	0.108	1.75	0.037	0.5	0.031	0.35
	ATP-GTCC	0.069	1.1	0.015	0.1	0.014	0.09
VSDC	MFCC-GTCC	0.49	7.33	–	–	–	–
	ATP-MFCC	0.089	1.33	–	–	–	–
	ATP-GTCC	0.04	0.6	–	–	–	–

Table 5
Detection performance of different classifiers with proposed features.

Dataset	Classifiers	Replay		Cloning		Cloned-Replay	
		min-tDCF	EER%	min-tDCF	EER%	min-tDCF	EER%
VSDC	Decision Trees	0.3713	16.33	–	–	–	–
	Naïve Bayes	0.528	27	–	–	–	–
	KNN	0.045	0.75	–	–	–	–
	Ensemble bagged trees	0.115	1.83	–	–	–	–
	BiLSTM	0.3132	13.1	–	–	–	–
	SVM	0.04	0.6	–	–	–	–
ASVspoof 2019	Decision Trees	0.361	15	0.112	5	0.05	0.75
	Naïve Bayes	0.426	19.75	0.098	2.8	0.072	1.41
	KNN	0.139	6.75	0.088	2	0.032	0.5
	Ensemble Models	0.149	7	0.03	0.4	0.048	0.75
	BiLSTM	0.323	12.7	0.091	2.2	0.007	0.1
	SVM	0.064	1	0.01	0.15	0.006	0.09

cloned-replay spoofing detection respectively. Therefore, we claim that SVM trained on our ATP-GTCC features can effectively be used to classify the bonafide and spoof audios.

We also observed that all machine learning classifiers achieved reasonably well for replay attacks detection, however, performs remarkably well for voice cloning and cloned replay detection. The main reason of this significant difference in performance between replay and cloning spoofing attacks is the lack of high-quality cloning audio samples in ASVspoof 2019 dataset. This also highlights the need to develop more high-quality cloning audios that can make this problem reasonably challenging.

3.2.5. Performance comparison with existing countermeasures

To evaluate the significance of the proposed countermeasure for voice spoofing detection, we compared it with existing state-of-the-art voice spoofing countermeasures [5,13,15,16,23]. The results obtained on the ASVspoof 2019 dataset are shown in Table 6. For PA set, [5] performs worst and [23] performs the best, whereas, the proposed method also achieved remarkable results with min-tDCF of 0.069 and EER of 1.1%. Similarly, for LA set, [13] performs worst, whereas, the proposed method performs best and achieved min-tDCF of 0.015 and EER of 0.1%. From this analysis, we can conclude that our spoofing countermeasure can reliably be used to detect a variety of voice spoofing attacks.

Table 6
Performance comparison with existing countermeasures for voice spoofing detection.

Spoofing Type	Features	min-tDCF	EER%
PA	Yang et al. [5] (CMOC/A-DNN)	0.208	11.447
	Yang et al. [5] CVOC/A-DNN	0.178	9.269
	Monteiro et al. [15] (LFCC + ProdSpec + MGDCC–CNN)	0.07	2.015
	Lavrentyeva et al. [16] (CQT + LFCC + DCT-LCNN)	0.0122	0.54
	Yamagishi et al. [22] (CQCC-GMM baseline)	0.2454	11.04
	Yamagishi et al. [22] (LFCC-GMM baseline)	0.3017	13.54
	Todisco et al. [23] (Deep Features)	0.0096	0.39
	Proposed Method (ATP + GTCC-SVM)	0.069	1.1
LA	Yang et al. [13] (ICQC + ICQCC + ICBC + ICLBC-DNN)	0.237	10.44
	Monteiro et al. [15] (LFCC + ProdSpec + MGDCC–CNN)	0.198	9.09
	Lavrentyeva et al. [16] (CQT + LFCC + DCT-LCNN)	0.051	1.84
	Yamagishi et al. [22] (CQCC-GMM baseline)	0.236	9.87
	Yamagishi et al. [22] (LFCC-GMM baseline)	0.212	11.96
	Todisco et al. [23] (Deep Features)	0.0069	0.22
	Proposed Method (ATP + GTCC-SVM)	0.015	0.1

3.2.6. Performance comparison with existing features

To evaluate the significance of ATP and our fused ATP-GTCC feature descriptor, we compared it against the comparative features for voice spoofing detection. We selected those features that was used for both PA and LA attacks detection. More specifically, we compared our proposed features against the baseline (CQCC and LFCC), and CQT-LFCC-DCT features using the SVM classifier. Performance obtained on the ASVspoof 2019 dataset is presented in Table 7. The proposed features achieved the best results by obtaining min-tDCF and EER of 0.069 and 1.1% for PA, and 0.015 and 0.1% for LA collection of ASVspoof 2019 dataset over other features. LFCC features perform second best and achieves the min-tDCF and EER of 0.762 and 29.44% for PA, and 0.769 and 29.75% for LA collection. Whereas, CQCC performed the worst and achieved the highest min-tDCF and EER of 0.812 and 36.82% for PA, and 0.815 and 36.98% for LA. From these results, we can conclude that our proposed features outperform the existing state-of-the-art features for voice spoofing detection. These results indicate the significance of our ATP-GTCC features for effectively capturing the dynamic speaker induced variations in bonafide signal, algorithmic artifacts in cloning algorithm and microphone distortions in the replay signal.

3.3. Discussion

To develop a robust method to detect various voice spoofing/presentation attacks is an important requirement for applications of countermeasure/presentation attack detectors and automated speaker recognition systems. Literature shows that countermeasures trained with one class of spoofing attacks fail to generalize well for other classes of spoofing attacks [25,26]. For example, systems trained with speech cloning show poor performance for replay detection [27]. The findings of the first two ASVspoof challenges also reveal that the playback voice recording in a new replay session is difficult to detect [3]. To address this important problem, this paper lay the foundation for developing spoofing detector to

Table 7
Performance comparison with existing features on SVM for voice spoofing detection.

Spoofing Type	Features	min-tDCF	EER%
PA	CQT-LFCC-DCT [16]	0.748	28.19
	CQCC baseline [22]	0.982	36.82
	LFCC baseline [22]	0.762	29.44
	Proposed ATP-GTCC	0.069	1.1
LA	CQT-LFCC-DCT [16]	0.751	28.85
	CQCC baseline [22]	0.985	36.98
	LFCC baseline [22]	0.769	29.75
	Proposed ATP-GTCC	0.015	0.1

detect speech synthesis, single- and multiorder-playback, and cloned-replay attacks for VCS by exploiting novel features descriptor i.e. Acoustic Ternary Pattern and fused with GTCC features. The human speech contains dynamic speaker induced variations in comparison to synthetic voice. For example, the natural pauses of human model of speech production are missing from the synthetic voice generated by deep learning based speech cloning algorithms [24]. The human model of speech production has 8,000 to 40,000 data points per second. On the contrary, synthetic voice sounds similar and does have very low standard deviation in terms of data points compared to human models. From the results (Table 2) of our first experiment (*Section 3.2.1*), we can clearly observe that ATP features achieve remarkable results in terms of detecting the replay, cloning, and cloned-replay attack. Specifically, ATP performs best for voice cloning detection which proves our hypothesis that ATP has the capability to accurately capture the dynamic attributes of human's speech variations that are absent in the synthetic speech. Further, generative approaches, such as Wavenet and Tacotron, find it hard to differentiate between general noise and speech in a training dataset resulting in noise packaged in as part of the cloned voice. From the results it is clear when ATP is combined with GTCC, the overall detection performance is further increased as this robotic 'whine' is nicely captured by GTCC. From the results (Table 2), we can observe that the ATP-GTCC features provide superior detection performance for replay detection on both the VSDC and ASVspoof datasets. Our results also prove our second hypothesis that the microphone induced distortions in the replay/-playback samples can be effectively captured through ATP-GTCC features. These distortions become further amplified when same voice is played back in the chain, thus our model more accurately captures *n*th-order replay samples. It is important to mention that we have also introduced a *novel voice spoofing attack (cloned-replay)* which is unknown to research community. Since the resultant signal of cloned replay contains properties of both replay and cloned voice, thus ATP-GTCC accurately identifies patterns of cloned replay attack. Performance evaluation on a variety of voice spoofing attacks using two publicly available and diverse datasets (i.e. ASVspoof 2019 and our own VSDC) signify the effectiveness of our unified method for voice spoofing detection in VCS.

4. Conclusion

The proposed voice spoofing countermeasure is the first attempt to address the issue of multi-order voice spoofing attacks, synthetic voice attack, and cloned replay attacks. We proposed ATP-GTCC features to effectively capture the distortions of 1st- and 2nd-order spoofing samples, absence of the dynamic attributes of human's speech variations in synthetic voice and presence of robotic noise in it, and cloned replay attacks. The proposed model was evaluated on samples of VSDC and ASV2019 datasets which were recorded under variety of environmental conditions, diverse recording and play back equipment, and various state-of-the-art deep learning based generative models. Additionally, we introduced a novel voice spoofing threat i.e. *cloned-replay* that can also be used to spoof the VCSs. We demonstrated through experiments that our proposed features can reliably be used to detect multiple voice spoofing attacks. In the future, we plan to improve our countermeasure more robust on cross dataset scenario.

CRediT authorship contribution statement

Ali Javed: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing - original draft, Visualization, Experimental design, Writing - review and editing. **Khalid Mahmood Malik:** Conceptualization, Formal Analysis, Investigation, Project Administration, Resources, Methodology, Data Curation, Writing - original draft, Experimental design, Supervision, Writing - review and editing, Funding Acquisition. **Aun Irtaza:** Investigation, Methodology, Software, Experimental design, Writing - review and editing. **Hafiz Malik:** Resources, Methodology, Experimental design, Supervision, Writing - review and editing, Funding Acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by grant of National Science Foundation of USA via Awards No. (1815724) and (1816019).

References

[1] Harvel D. An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. [online]. Available: https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/, 2019.
[2] Malik KM, Malik H, Baumann R. sTowards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). p. 523–8.
[3] Sahidullah M, Delgado H, Todisco M, Kinnunen T, Evans N, Yamagishi J, Lee KA Introduction to voice presentation attack detection and recent advances. In Handbook of Biometric Anti-Spoofing, 2019, pp. 321–361. Springer, Cham..
[4] Todisco M, Delgado H, Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Comput Speech Language 2017;45:516–35.
[5] Yang J, Xu L, Ren B, Ji Y. Discriminative features based on modified log magnitude spectrum for playback speech detection. EURASIP J Audio Speech Music Process 2020;2020:1–14.
[6] Gunendradasan T, Irtza S, Ambikairajah E, Epps J. Transmission line cochlear model based am-fm features for replay attack detection. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 6136–40.
[7] Witkowski M, Kacprzak S, Zelasko P, Kowalczyk K, Galka J. Audio Replay Attack Detection Using High-Frequency Features. In Interspeech (pp. 27–31), 2017.
[8] Chen Z, Xie Z, Zhang W, Xu X. ResNet and Model Fusion for Automatic Spoofing Detection, in INTERSPEECH, 2017 (pp. 102–106).
[9] Chettri B, Kinnunen T, Benetos E. Deep generative variational autoencoding for replay spoof detection in automatic speaker verification. Comput Speech Language 2020;101092.
[10] De Leon PL, Pucher M, Yamagishi J, Hernaez I, Saratxaga I. Evaluation of speaker verification security and detection of HMM-based synthetic speech. IEEE Trans Audio Speech Language Process 2012;20(8):2280–90.
[11] Saratxaga I, Sanchez J, Wu Z, Hernaez I, Navas E. Synthetic speech detection using phase information. Speech Commun 2016;81:30–41.
[12] Patel TB, Patil HA. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In: Sixteenth Annual Conference of the International Speech Communication Association.
[13] Yang J, Das RK. Long-term high frequency features for synthetic speech detection. Digital Signal Process 2020;97: 102622.
[14] Malik H. Securing Voice-driven Interfaces against Fake (Cloned) Audio Attacks. In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). p. 512–7.
[15] Monteiro J, Alam J, Falk TH. Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. Comput Speech Language 2020;101096.
[16] Lavrentyeva G, Novoselov S, Tseren A, Volkova M, Gorlanov A, Kozlov A. Stc antispoofing systems for the asvspoof2019 challenge. arXiv preprint arXiv:1904.05576, 2019.
[17] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 1996;29 (1):51–9.
[18] Valero X, Alias F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Trans Multimedia 2012;14 (6):1684–9.
[19] Saranya MS, Padmanabhan R, Murthy HA. Replay attack detection in speaker verification using non-voiced segments and decision level feature switching. In: 2018 International Conference on Signal Processing and Communications (SPCOM). IEEE; 2018. p. 332–6.
[20] Javed A. ATP-GTCC Features [Online]. Available: https://github.com/alijaved21/ATP-GTCC-Features.

[21] Baumann R, Malik KM, Javed A, Ball A, Kujawa B, Malik H. Voice Spoofing Detection Corpus for Single and Multi-order Audio Replays. Comput Speech Language 2021;65: 101132.

[22] Yamagishi J, Todisco M, Sahidullah M, Delgado H, Wang X, Evans N., et al. ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database, 2019.

[23] Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, et al. Asvspoof 2019: Future horizons in spoofed and fake audio detection. arXiv preprint arXiv:1904.05441, 2019.

[24] Mwiti D, A 2019 Guide to Speech Synthesis with Deep Learning. [online]. Available: https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd#70c3, 2019 .

[25] Korshunov P, Marcel S. Cross-database evaluation of audio-based spoofing detection systems. In: Interspeech (No. CONF), 2016.

[26] Gonçalves AR, Violato RP, Korshunov P, Marcel S, Simoes FO. On the generalization of fused systems in voice presentation attack detection. In: 2017 International conference of the biometrics special interest group (BIOSIG). IEEE; 2017. p. 1–5.

[27] Paul D, Sahidullah M, Saha G. Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2017. p. 2047–51.