Odyssey 2018 The Speaker and Language Recognition Workshop
26-29 June 2018, Les Sables d'Olonne, France

A Spoofing Benchmark for the 2018 Voice Conversion Challenge: Leveraging from Spoofing Countermeasures for Speech Artifact Assessment

Tomi Kinnunen[1], *Jaime Lorenzo-Trueba*[2], *Junichi Yamagishi*[2,3], *Tomoki Toda*[4], *Daisuke Saito*[5], *Fernando Villavicencio*[6], *Zhenhua Ling*[7]

[1] University of Eastern Finland, Joensuu, Finland
[2] National Institute of Informatics, Tokyo, Japan [3] University of Edinburgh, UK
[4] Nagoya University, Nagoya, Japan [5] University of Tokyo, Tokyo, Japan
[6] ObEN, Pasadena, USA [7] University of Science and Technology of China, Heifei, China
vcc2018@vc-challenge.org

Abstract

Voice conversion (VC) aims at conversion of speaker characteristic without altering content. Due to training data limitations and modeling imperfections, it is difficult to achieve believable speaker mimicry without introducing processing artifacts; performance assessment of VC, therefore, usually involves both speaker similarity and quality evaluation by a human panel. As a time-consuming, expensive, and non-reproducible process, it hinders rapid prototyping of new VC technology. We address artifact assessment using an alternative, objective approach leveraging from prior work on spoofing countermeasures (CMs) for automatic speaker verification. Therein, CMs are used for rejecting 'fake' inputs such as replayed, synthetic or converted speech but their potential for automatic speech artifact assessment remains unknown. This study serves to fill that gap. As a supplement to subjective results for the 2018 Voice Conversion Challenge (VCC'18) data, we configure a standard constant-Q cepstral coefficient CM to quantify the extent of processing artifacts. Equal error rate (EER) of the CM, a confusability index of VC samples with real human speech, serves as our artifact measure. Two clusters of VCC'18 entries are identified: low-quality ones with detectable artifacts (low EERs), and higher quality ones with less artifacts. None of the VCC'18 systems, however, is perfect: all EERs are $< 30\%$ (the 'ideal' value would be 50%). Our preliminary findings suggest potential of CMs outside of their original application, as a supplemental optimization and benchmarking tool to enhance VC technology.

1. Introduction

Voice conversion (VC) [1, 2] aims at conversion of speaker characteristic without altering the speech content. Typical use of VC technology includes applications in entertainment industry, such as customizing artificial voices for audio-books and games. In such applications, the VC samples are optimized for human listeners. In the recent past, thanks to technological advances in both VC and automatic speaker verification (ASV) technology, VC finds also frequent use in assessing ASV system vulnerability against intentional circumvention (spoofing) [3]. In this case, the VC samples are prepared for the ASV system. In both cases, the goal is that the VC samples manage to make the primary observer (either a human or a machine) to believe they are observing a certain targeted speaker that is different from the source speaker. But human perception and machine perception are different, and in the case of ASV and its spoofing, machine perception is more relevant.

Even if the VC technology itself has evolved steadily over the years [4, 5, 6], the evaluation methods of VC are more varied compared to tasks such as ASV or automatic speech recognition. The primary evaluation methods are perceptual tests since the target of the above applications is normally human perception. In addition, log-spectral distortion and cepstral distortion (between converted and target utterances) are also used as supplementary information. At this moment, we lack of universally adopted objective performance measures. Furthermore, there was no standard database until recently.

Given the situation, [7] launched a *Voice Conversion Challenge* (VCC) series in 2016, with a follow-up in 2018[1] organized by the authors of this study [8]. The primary methodology of the evaluation of VC systems, including VCC, is a perceptual test of various VC systems trained on a common corpus. The perceptual test usually involves both speaker similarity and quality evaluation by a human panel since it is difficult to achieve convincing speaker transformation without introducing processing artifacts due to training data limitations and modeling imperfections. The VCC series provide results of large-scale perceptual tests that compare many different types of VC systems on a comparable basis. This is helpful towards understanding human perception and optimization strategies employed by listeners. On the other hand, listening test results do not directly reflect spoofing capability and we do not know how the results of the perceptual test are related to machine perceptions, that is, ASV and its spoofing.

With the above motivations in mind, the present study accompanies [8] that provides details of the 2018 challenge data, analysis of the submitted systems and the perceptual results. We provide supplemental objective quality results on the degrees of artifacts that each of the submitted systems have. Even if our experiments are framed to the context of the latest VCC'18 challenge, our contribution is that of a novel objective speech artifact assessment leveraging from the rapidly emerging topic of *spoofing countermeasures*. In the context of ASV, *spoofing* refers to intentional circumvention of the ASV system to obtain illegitimate access as another targeted user [9] — VC technology being a representative example. ASV vulnerability due to spoofing has been known for about two decades [10] but has gained momentum only relatively recently with increased interest towards ASV deployment for user authentication, as well as

[1] http://www.vc-challenge.org/, data available at http://dx.doi.org/10.7488/ds/2337 since April 10, 2018.

10.21437/Odyssey.2018-27

availability of common evaluation resources [11, 12] to enable meaningful comparisons of different spoofing countermeasures.

There has been continued research towards generalized spoofing countermeasures to detect spoofing attacks more accurately. As a result, several advanced front-end [13, 14, 15, 16] and machine learning [17] oriented techniques have been developed for the task of detecting the presence of a spoofing attack in a given audio segment. The task is framed as a hypothesis testing problem with *bona fide* (legitimate human speech) hypothesis as the null hypothesis and *spoof* as an alternative hypothesis. The exact definition of the latter depends on the type of the spoofing attack (*e.g.* VC or replay attack).

If the detector is carefully optimized and the probability distributions of bona fide and spoof classes are sufficiently distinct, one can detect the attacks. But if the spoofed samples resemble too closely real human speech, the detector can mistakenly classify them as bona fide speech. Therefore, the number of errors made by a spoofing countermeasure for a given batch of test files is associated with how closely the spoof samples resemble the bona fide samples. In specific, the spoofing countermeasure gauges the amount of speech artifacts that only the spoof samples have (regardless whether the artifacts are audible to a human or not) and tell us how close the spoof samples are compared to the bona fide samples. Therefore we hypothesize that this is useful for the automatic assessment of speech artifacts produced by VC process and may be used as one of objective performance measures. Therefore, this study compares the performance of the spoofing countermeasure with subjective quality evaluation results obtained in VCC'18 and investigate how they are related to each other.

2. Subjective quality of converted voices

Suppose we have a patch of N source speaker utterances $\mathbf{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_N\}$ processed through S voice conversion systems $s = 1, \ldots, S$. We denote the utterances converted by system s by $\mathbf{Y}^s = \{\mathcal{Y}_1^s, \ldots, \mathcal{Y}_N^s\}$ and use $\mathbf{Y} = \cup_{s=1}^{S} \mathbf{Y}^s$ to denote all the $N \cdot S$ conversions. For the purpose of comparing the alternative VC systems in terms of speech quality, the samples $\mathbf{Y}$ are collectively listened to by a cohort of human *observers*, $\mathcal{O} = \{O_1, \ldots, O_L\}$, each of who outputs an *opinion score* for a subset of samples indexed by α_i for observer i. Note that not all the observers necessarily listen to the same utterances, nor necessarily listen to even the same number of samples. The process of obtaining the opinion score of observer i for utterance $\mathcal{Y} \in \mathbf{Y}$ can be thought of as evaluating an abstract, non-deterministic and possibly time-varying, function $h(\mathcal{Y}, O_i)$, that models human listening mechanism of O_i. It depends on many factors such as the listener's life experience and concentration, the listening environment, audio equipment used, and familiarity with the language. We do not have access to the internals of $h(\cdot, O_i)$ but only its observed output, in this study the standard 5-point rating scale ranging from 1 (lowest quality) to 5 (highest quality).

Because of random variation in the outputs, caused by differences in $h(\cdot, O_i)$, one represents the results of the listening panel in an averaged form. The well-known population summary measure, *mean opinion score* (MOS), is computed for system s by $\text{MOS}_s = (1/L_s) \sum_{n=1}^{N} \sum_{i=1}^{L} h(\mathcal{Y}_n^s, O_i)$, where L_s is the total number of opinion scores obtained for the samples of system s, and where we assign a dummy value $h(\mathcal{Y}, O_i) = 0$ if listener i did not rate $\mathcal{Y}$. The higher the MOS value, the higher quality the samples of system s. The definition of 'quality' is also subjective. No instruction on what high-quality or low-quality means is normally given.

3. Proposed objective artifact assessment using spoofing countermeasures

Here we want to construct a model that automatically scores the amount of speech artifacts. The artifacts may be audible or non-audible[2].

3.1. Obtaining machine scores

With the objective artifact estimation, the problem setup is the same as above: given $\mathbf{Y}^s$, we want to obtain a single numerical value, similar to MOS, that relates to the degree of artifacts produced by a VC system s. To this end, we replace the abstract human observer $h(\mathcal{Y}, \mathcal{O}_i)$ by a *machine* observer, $m(\mathcal{Y}, \boldsymbol{\theta})$, represented by some model parameters $\boldsymbol{\theta}$. There are several differences between h and m. First, unlike h, evaluating m is *deterministic* and *time-invariant* — in other words, it yields the same output when repeated on sample $\mathcal{X}$, and is not dependent on the time when invoked. Second, unlike h where one usually constrains the outputs to be quantized to a small set of ordinal variables, we allow the domain of m to be the entire real line $\mathbb{R}$; the scale of the output value is arbitrary, but similar to h, higher numerical values in relative terms indicate higher speech quality as judged by the observer $\boldsymbol{\theta}$.

To flesh out the above vague idea, in this work $m(\cdot, \boldsymbol{\theta})$ takes the form of a *likelihood ratio detector*. Likelihood ratios arise naturally from the Bayes theorem and serve as the starting point for making statistically optimal decisions. For a given input utterance $\mathcal{Y} \in \mathbf{Y}$, we compute a *log-likelihood ratio* (LLR) score,

$$\ell(\mathcal{Y}|\boldsymbol{\theta}) = \log \frac{p(\mathcal{Y}|H_0)}{p(\mathcal{Y}|H_1)} = \log \frac{p(\mathcal{Y}|\boldsymbol{\theta}_{\text{nat}})}{p(\mathcal{Y}|\boldsymbol{\theta}_{\text{artif}})}, \quad (1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{nat}}, \boldsymbol{\theta}_{\text{artif}})$. The null hypothesis H_0, modeled through $\boldsymbol{\theta}_{\text{nat}}$, is that $\mathcal{Y}$ represents natural human speech without artifacts. The alternative hypothesis H_1, modeled using $\boldsymbol{\theta}_{\text{artif}}$, states that $\mathcal{Y}$ originates from artificial speech generation (such as voice conversion or speech synthesis). Higher numerical values are therefore associated with speech that appears more 'human-like', lacking vocoding artifacts or other problems that VC systems tend to generate.

To train $\boldsymbol{\theta}_{\text{nat}}$, we gather a large collection of natural human utterances and train the model from the pooled data; similarly, to train $\boldsymbol{\theta}_{\text{artif}}$, we gather a representative collection of artificial speech samples (such as samples from several state-of-the-art VC systems, *prior* to evaluating a new VC system). In this work, we use Gaussian mixture model (GMM) to model each hypothesis, trained through expectation-maximization (EM) algorithm. Though more advanced spoofing countermeasure backends are available, GMMs produced good results for the ASVspoof'15 challenge consisting also of high-quality clean samples as here, with the benefit of simplicity.

3.2. Error rate of the CM as an objective quality measure

The log-likelihood ratio in (1) outputs a number for a single utterance $\mathcal{Y}$. Now, how do we obtain a summary value for all the samples of a given VC system s? While it might be appealing to just average $\ell(\mathcal{Y}_n|\boldsymbol{\theta})$ over the samples $\mathcal{Y}_n \in \mathbf{Y}^s$, similar to MOS, this is not recommendable. Unlike the opinion score, ℓ is

[2]Hence the aim of the measure is not to approximate the subjective quality judgement.

unbounded and is therefore more difficult to interpret. Further, the model $\boldsymbol{\theta}$ is imperfect version of reality and cannot possibly produce meaningful LLR scores for all human speech and all spoofing attacks. As a result, the scale of ℓ is arbitrary and dependent on the various modeling choices (including that of feature extraction). The LLRs across databases, VC systems and converted utterances are not in a commensurable scale. In the nomenclature of ASV literature, we might say ℓ is not *well-calibrated* [18].

Hence, it is better to adopt a summary measure that is more intuitive and comparable across different evaluation environments. Our proposal is the *error rate* of the detector, as a measure of its ability to differentiate authentic human utterances from those generated by voice conversion. Our philosophy is as follows. If the samples from a VC system S_1 manage to fool a given artificial speech detector more often than samples from another competitive VC system S_2, we can say samples of S_1 appear more human-like in the eyes[3] of the artificial speech detector. Utterances having less processing artifacts are more difficult to discriminative from human speech, giving rise to higher error rate of the detector.

Even if our inspiration for such a proposal originates from the work in ASV anti-spoofing, the viewpoint is now switched from *defender* to *attacker*. In the ASV anti-spoofing, one keeps improving spoofing countermeasures so that they are more accurate in detecting advanced speech synthesis and voice conversion attacks (the lower the error rate, the better). But now we consider the anti-spoofing system to be fixed with the goal of improving the performance of voice conversion systems — the higher the error rate, the better.

As for the actual error rate measurement, we adopt the standard metric of *equal error rate* (EER) used extensively in ASV, anti-spoofing and biometrics research. The output scores (estimated LLRs) produce two different types of errors, *false alarms* and *misses*, that are traded off with respect to each other. Here, false alarm (false acceptance) rate (FAR) is the proportion of artificial speech samples that the detector falsely accepted as *bona fide* (human) samples. Miss (or false rejection) rate (MR), in turn, is the proportion of falsely rejected bona fide samples. FAR and MR are, respectively, decreasing and increasing functions of a detection threshold. The EER, then, is the unique error rate corresponding to the threshold at which FAR and MR equal each other. As the detection task involves only two classes, the chance level is EER of 50%. This would be our 'ideal' value for a successful artifact-free VC system[4]. One might therefore optionally report a scaled version $\frac{1}{10}$EER %, to give rise to a continuous version of a 5-point 'opinion score', judged by the machine observer.

3.3. Choice of the countermeasure model (1)

To implement the artifact LLR detector of (1), we represent speech utterances using a sequence of short-term spectral features, $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ with $\boldsymbol{y}_n \in \mathbb{R}^D$, D being the feature dimensionality. At the training stage, we use the pooled feature vectors from each class to independently train two Gaussian mixture models (GMMs), $\boldsymbol{\theta}_{\text{nat}}$ and $\boldsymbol{\theta}_{\text{artif}}$, using the standard *expectation-maximization* (EM) algorithm. We use diagonal covariance matrices and consider the number of Gaussian components, C, as a tuning parameter. It can be used to adjust the balance between over- and under-fitting.

Choice of the short-term feature representation is critically important in the context of spoofing countermeasures [13]. Findings from the first ASVspoof 2015 challenge [11] highlighted the importance of spectral and temporal *details* for the task of discriminating real and spoofed speech. In specific, conventional MFCCs, a *low-resolution*, *low-frequency focusing* feature set that has enjoyed its *de facto* audio representation status for almost 40 years, is a suboptimal choice. Discriminating human speech from synthetic or converted speech and identifying the artifacts seems to require more detailed time-frequency representation. The winning system of the ASVspoof'15 challenge [16] used MFCCs with a combination of cochlear filter cepstral coefficients and instantaneous frequency. Later, [14] introduced a single feature set, *constant-Q cepstral coefficients* (CQCCs), based on the constant-Q transform [19]. It lead to the lowest reported EERs on the ASVspoof'15 corpus at the time. Substantial follow-up work (*e.g.* [15]) has improved feature extractors even further. In this work we use CQCCs due to their high reported detection accuracy, simplicity, and widespread adoption by the research community. We use an open-source CQCC implementation provided to the second ASVspoof challenge participants[5], and similarly another public toolkit to train the GMMs [20].

3.4. Data-related considerations to enable fair evaluation

Besides specification of the front-end features, another key consideration is the choice of training and development data of the countermeasure. Despite high accuracy of the spoofing countermeasure front-ends listed above, they are notoriously sensitive to training-test mismatch (cross-corpus performance) [21], additive noise [22] and channel/bandwidth mismatch [23]. In short, countermeasures are easy to overfit for specific data leading to potentially arbitrarily bad results for a different test data. This makes the selection of both data, and optimization process of the countermeasure parameters important.

As challenge organizers, our responsibility is to avoid favoring any specific VC system but provide as unbiased assessment of the systems as we possibly can. To this end, we aim to optimize our countermeasure with the following requirements in mind.

1. **Stability across datasets.** The countermeasure should show stable enough results when executed on different corpora so that one can trust the result to be less dependent on the specifics of VCC'18 data.
2. **Detection of state-of-the-art voice conversion.** We should aim at detecting the current state-of-the-art, or otherwise previously known, VC attacks, with sufficient accuracy.
3. **No tweaking using participant submissions.** We should *not* optimize our proposed measure with feedback of the VCC'18 evaluation entries. That is, we should not look at the error rates and use the submitted samples to enrich training sets. Instead, we should fix the data and parameters to the best-known values using *data that precedes VCC'18 participant entries*.

The last requirement might be less obvious to the reader from the voice conversion field. It reflects the viewpoint of

[3] Or ears?

[4] Technically speaking it is possible to do *worse* than the coin-flipping rate, for instance by swapping the two model likelihoods in (1). EERs much larger than 50 % suggest usually an implementation bugs of the detector, and are not interesting from the perspective of evaluation.

[5] http://www.asvspoof.org/data2017/baseline_CM.zip

Table 1: Considered datasets for the construction of spoofing countermeasures for objective quality assessment of the VCC'18 samples. The datasets differ in the number of speakers and diversity of spoofing attacks. The ASVspoof'15 data represented state-of-the-art of SS and VC in 2014-2015, while VCC'16 contains more modern attacks. **ASVspoof:** *Automatic Speaker Verification Spoofing and Countermeasures Challenge*, **VCC:** *Voice Conversion Challenge*, **SS:** *speech synthesis*, **VC:** *voice conversion*, **STRAIGHT**: *Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum*, and **LPC**: *linear predictive coding*.

	ASVspoof'15 train	**ASVspoof'15 dev**	**ASVspoof'15 eval**	**VCC'16**	**VCC'18 base**
Types of attacks	SS and VC	SS and VC	SS and VC	VC	VC
Waveform generation	STRAIGHT, MLSA	STRAIGHT	STRAIGHT, diphone concatenation	STRAIGHT, LPC, Ahocoder	Waveform filtering
# Spoofing attacks	5	10	10	18	1 (VCC'18 basel.)
# Spoof files	12,625	49,875	184,000	24,300	2,240
# Speakers (♂+ ♀)	25 (10 + 15)	35 (15 + 20)	46 (20 + 26)	10 (5+5)	12 (6 + 6)
# Human files	3,750	3,497	9,404	535	1,139

the spoofing countermeasure as a security gate in real-world deployment: one does not know the attacks (here, voice conversion samples) in advance but has to use her best knowledge to prepare the countermeasure using attacks available beforehand. In the standard automatic speaker verification (ASV) evaluation benchmarks conducted by National Institute of Standards and Technology (NIST), evaluation participants are similarly expected to process the new evaluation samples completely blindly — they are not allowed to interact with the evaluation samples in any manner (such as by listening to them, or using them to do modeling decisions), for the same reason. We apply the same principle in our role as a challenge evaluator, to give all the submitted VCC'18 systems an equal opportunity to break down our countermeasure.

4. Experimental data

4.1. The voice conversion challenge 2018

The 2018 Voice Conversion Challenge (VCC'18) is a follow-up to the VCC series kicked off in 2016 [7]. It features the task of speaker identity conversion. A detailed description of the challenge data, rules, analysis of the submitted systems and extensive perceptual results are provided in another paper [8], with only the key facts repeated here for completeness.

The VCC'18 challenge is designed to promote development of both *parallel* and *nonparallel* VC methods; the parallel (**Hub**) task contains source-target training utterances with matched speech content while in the nonparallel (**Spoke**) task, the contents differ. Both tasks contain the same target speaker data but the source speakers are different. The Hub task formed the core (required) task for the registrants, whereas participation to the Spoke task was optional. The participants were provided with training and development data, and were asked to submit their converted audio files for previously unseen source utterances.

Both the VCC'16 and the VCC'18 data are based on the DAPS (Data And Production Speech) dataset [24] including native US English speakers recording in a professional setting. The source and the target speakers across the two challenges are all disjoint. The number of test sentences is 35 and the participants were asked to submit the converted voices for a total of 16 source-target speaker pairs in both Spoke and Hub tasks. The results were evaluated subjectively using crowd-sourcing. A total of 267 unique crowdworkers evaluated perceptually both naturalness and speaker similarity. The former ranged from 1 (completely unnatural) to 5 (completely natural) while a 4-point scale was used for the latter ("Same, absolutely sure", "Same, not sure", "Different, not sure", "Different, absolutely sure". The trials consisted of comparisons of VC samples with either the source speaker or the target speaker.

4.2. Data for countermeasure development

With the above considerations in mind, we involve data from several audio collection that contain both synthetic speech and converted voice samples. The datasets are summarized in Table 2. As for the **ASVspoof 2015** collection (train, dev, eval), documented in detail in [11], we follow the protocol files provided with the corpus[6]. The **VCC'16** data, in turn, consists of the samples of the first edition of the VCC series [7]. In specific, it contains the participant submissions from 17 different systems plus 1 baseline system, along with samples from 10 speakers (three source males, two source males, two target females, three target males).

The last data, denoted **VCC'18 base**, contains *only* the VCC'18 baseline as its only attack (in specific, samples of the VCC'18 baseline system). The human trials are the same as those in VCC'16 augmented with two of the VCC'18 speakers. Again no submitted VCC'18 system is used for training the countermeasures.

5. Results

5.1. CQCC-GMM countermeasure optimization

Given the sampling rate mismatches across the prior corpora (16 kHz for ASVspoof'15 and VCC'16) and the new VCC'18 data (22.05 kHz), we downsample the latter to 16 kHz. In our first experiment, we study the performance of the CQCC-GMM detector for different selections of training and test data, with the primary goal being selection of our main training data. For this first experiment, we fix the CQCC configuration to the default setting of the ASVspoof'17 challenge baseline (29 base CQCCs plus the zeroth (energy) coefficient, with deltas and double deltas, giving 90-dimensional features. We study the CQCC configurations both without any feature normalization, as well as with cepstral mean and variance normalization (CMVN); it was included since it might be potentially helpful in

[6] https://datashare.is.ed.ac.uk/handle/10283/853

Table 2: Equal error rate (EER, %) for intra-corpus (ASVspoof'15) and cross-corpus artifact detection experiments, the *lower* the better. Number of Gaussians 32, 29 CQCCs and energy coefficient with deltas and double deltas. The EERs are averages of attack-specific EERs.

	Train	Test	CQCC (raw)	CQCC (CMVN)
(i)	TRAIN'15	DEV'15	0.38	1.99
(ii)	TRAIN'15	EVAL'15	1.83	2.21
(iii)	TRAIN'15	VCC'16	35.04	34.26
(iv)	ALL'15	VCC'16	34.86	31.52
(v)	VCC'16	DEV'15	26.18	33.34
(vi)	VCC'16	EVAL'15	21.48	34.75

Table 3: Equal error rate (EER, %) for CQCC optimization, the *lower* the better. Number of Gaussians 32, training data VCC'16. Lines (i) to (vii) are based on 29 CQCCs without the energy coefficient while (viii) contains the zeroth coefficient. The EERs are averages of attack-specific EERs.

		DEV'15		**VCC'18 base**	
	Front-end	CQCC (raw)	CQCC (CMVN)	CQCC (raw)	CQCC (CMVN)
(i)	stat	46.14	32.33	18.41	21.04
(ii)	Δ	12.11	31.20	24.85	21.55
(iii)	Δ^2	13.58	29.39	22.89	20.83
(iv)	Δ, Δ^2	**7.73**	**18.93**	22.08	16.94
(v)	stat, Δ	39.73	36.86	15.45	20.74
(vi)	stat, Δ^2	34.46	31.65	**13.83**	18.48
(vii)	stat, Δ, Δ^2	32.09	34.28	14.00	18.35
(viii)	z, stat, Δ, Δ^2	26.18	33.34	14.13	**16.74**

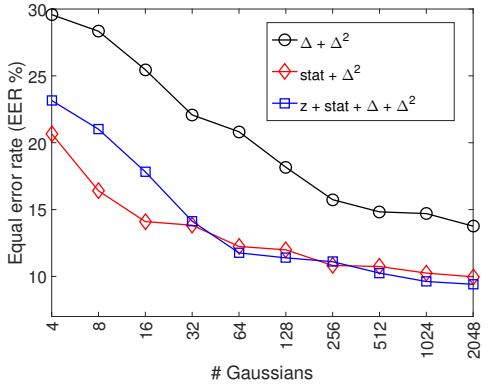


Figure 1: The results on VCC'18 base data with varied complexity of the GMM backend. Training data is VCC'16.

suppressing convolutive mismatch across datasets. Convolutive bias could originate from speaker, recording media or vocoder differences, and might be reducible through feature normalization techniques. In our implementation, we use utterance-level CMVN to obtain zero mean, unit variance features per file. We do not apply speech activity detection. The number of Gaussian components is set to 32.

The results for the ASVspoof'15 and VCC'16 data are shown in Table 2. The first two rows correspond to the standard protocols of the ASVspoof'15 corpus and reflect intra-corpus performance. The results are in line with the published literature. First, the error rates are remarkably low, demonstrating the potential of the CQCC-GMM countermeasure. Second, the error rate of the evaluation part is higher, due to presence of one *unknown* attack (S10). CMVN systematically degrades performance.

The last four rows of Table 2 show the cross-corpus performance. As expected, the error rates are now far higher. The training set ALL'15 indicated in line (iv) was obtained by pooling the train, dev and eval files of ASVspoof'15 into a large training set. Comparing experiments (iii) and (iv), the larger training gives no substantial boost for the unnormalized features (relative decrease of 0.5% in EER) though with some improvement (8% relative decrease) for the normalized features. CMVN is helpful in (iv) only. Comparing experiments (v) and (vi) to (iii) and (iv) of the unnormalized features, the results are not symmetric regarding the roles of training and test corpus. Even if the VCC'16 data is much smaller than ASVspoof'15 in terms of speaker and file count, the voice conversion samples (attacks) are perhaps more diverse acoustically.

Based on the results of Table 2, for the remainder of the experiments we fix VCC'16 as our training data. The next experiment concerns the impact delta features that have been noted accurate in detecting vocoded speech. Table 3 shows the results for the ASVspoof'15 dev (the more *difficult* one from dev and eval), and VCC'18 base. The results are shown again for raw and CMVN-processed features.

For the ASVspoof'15 dev trials, the outstanding front-end consists of just deltas and double deltas without feature normalization. The results highlight usefulness of the dynamic features, and *un*usefulness of the static coefficients — the first line consisting of static coefficients only gives performance close to the chance level of 50% EER. Comparing experiment (ii) to (v), experiment (iii) to (vi) and experiment (iv) to line (vii), inclusion of the static coefficients systematically degrades performance. It is also noteworthy that while the plain delta features are degraded by CMVN, CMVN boosts the performance of the static coefficients in cases (i), (v) and (vi). This might be partly explained noting that both CMVN and deltas are helping in reducing convolutive mismatch. For the VCC'18 base results, however, we have the opposite finding: CMVN helps the delta variants and systematically degrade the variants that include static coefficients. The globally best setups for both trial sets are obtained without CMVN. Comparing the best configurations from each trial set, 7.73% for the ASVspoof'15 dev and 13.83% for VCC'18 base, the latter appears harder. This indicates that the state-of-the-art voice conversion attack (VCC'18 baseline) is challenging for the CQCC-GMM countermeasure.

In our last parameter tweaking experiment, we fine-tune the number of Gaussians (that was 32 until now for computational reasons) using just the VCC'18 base data (training with VCC'16). Based on the results of Table 3, we select three representative feature set-ups, the one that produced good results on the ASVspoof'15 dev data; the best set-up (stat + Δ^2) for the VCC'18 base data, plus the full configuration (z + stat + Δ + Δ^2). The results displayed in Fig. 1 indicate improved performance with larger number of Gaussians as expected. The performance might slightly be improved by increasing the number of Gaussians further; due to resource reasons, we stopped at 2048. Concerning the front-end set-up, the full configuration containing static, delta, double delta and the zeroth coefficients

Table 4: Equal error rates (EER %) of the CQCC-GMM spoofing countermeasure of the VCC'18 entries on the Hub task. Here "B01" denotes the VCC'18 baseline system. The two countermeasures considered use deltas and double deltas ($\Delta + \Delta^2$) of 29 CQCCs, and 29 CQCCs plus zeroth coefficient along with deltas and double deltas (All feat.). Training data VCC'16, number of Gaussians 2048. The *higher* the EER, the better the VC system in terms of quality (less processing artifacts). ?: information unavailable to the authors.

Sys.	Waveform generation	$\Delta + \Delta^2$	All feat.	**Sys.**	Waveform generation	$\Delta + \Delta^2$	All feat.
B01	Waveform filtering	18.18	13.90	N09	Ahocoder	0.84	6.34
D01	World direct wave mod.	4.87	9.07	N10	Wavenet	2.49	7.06
D02	World	27.02	31.41	N11	STRAIGHT	0.45	0.71
D03	STRAIGHT	0.97	7.24	N12	Waveform filtering	24.11	11.61
D04	World	8.44	6.73	N13	World	5.69	8.18
D05	World	2.27	7.85	N14	Waveform filtering	24.57	13.55
N03	?	3.12	7.57	N15	World	4.26	8.39
N04	World	2.08	7.37	N16	Ahocoder	0.71	5.32
N05	SuperVP	25.74	19.63	N17	Wavenet	8.02	7.01
N06	World	1.59	36.47	N18	Griffin-Lim	17.94	10.22
N07	World	2.31	7.88	N19	World	6.48	10.56
N08	Waveform filtering	21.73	12.03	N20	World	2.02	6.82

yields the best results.

5.2. Results for the VCC'18 samples

We now fix our countermeasure parameters to compare the VCC'18 submissions. From Table 3, we see that CMVN helps only in 5 (out of 16) cases so we decide to not include it. Based on Table 3, the $\Delta + \Delta^2$ is a reasonable choice of features: it yields a clearly outstanding result on the difficult cross-corpus experiment with ASVspoof'15 dev and, with optimized number of Gaussians (Fig. 1), works reasonable well also for the VCC'18 base data. Additionally, we include the full feature set-up consisting of base coefficients (including energy) with deltas and double-deltas, as this yielded the best overall results in the detection of the baseline samples. Based on Fig. 1, we fix the number of Gaussians to 2048.

The results are shown in Table 4 for both of the selected feature setups on the VCC'18 Hub task. For ease of interpretation, we show the waveform generation method of each submission entry. For a case of the $\Delta + \Delta^2$ configuration, waveform filtering, SuperVP and Griffin-Lim based waveform generation methods were judged as VC methods that have relatively less artifacts compared to STRAIGHT, World, and Ahocoder vocoders. This is reasonable since they were proposed for improving issues of minimum phase vocoders. One suprising result to everyone may be that althought N10 was evaluated as the best VC by human listeners (about 4.1 MOS score), our methods detected its artifacts easily and its EER is low as 2.4 %. The N10 does not use any deterministic vocoders as above, but, uses μ-law quantized waveforms instead [25]. The μ-law quantization may cause obvious artifacts (although they are non audible to human and hence they were well evaluated by human listeners).

One interesting exception to our expectations is system D02. For the $\Delta + \Delta^2$ configuration, the VC system D02 has achieved highest EER although they have used the known vocoder. According to the listening test [8], the D02 samples sound very similar to source speakers and perfectly dissimilar to target speaker. This suggests that D02 used little modification of source speaker's waveforms and hence has likely less processing artifacts.

We can also see that the EERs are sensitive to the choice of acoustic features used. Using the full feature configuration, which was optimized to detect the VCC'18 baseline samples, the waveform filtering, SuperVP and Griffin-Lim based waveform generation methods still have higher EERs, but, system N06 has also very high EER (which, curiously, changes its ranking completely). We suspect that the countermeasure that uses full feature configuration is overfit and incapable of generalizing beyond the training data as well as the countermeasure with the $\Delta + \Delta^2$ front-end. It would be important for us to develop stable and robust models. Having said that, we can clearly see that none of the VC achieves EERs close to the chance level (50%) regardless of the selected features, indicating that all the VC methods produced artifacts more or less and having plenty of margin for improving the quality of the samples.

Finally, Fig. 2 displays a scatter plot of the EER of the $\Delta + \Delta^2$ countermeasure against the mean opinion score (MOS) from the perceptual experiments detailed in [8]. We do not observe strong association between the two but they are complementary to each other. This is not entirely surprising remembering that our objective measure focuses on audible and non-audible artifacts, whereas listeners evaluated audible naturalness subjectively. As expected, human perception and machine perception are different.

6. Conclusion

We have proposed the use of spoofing countermeasure for objective artifact assessment of converted voices as a supplement to the VCC'18 challenge results. Our approach is reference-free and text-independent in the sense that it does not require access neither to the original source speaker waveform nor any text transcripts. It assigns a single EER number to a batch of converted speech utterances from a single VC system, and can therefore be used in comparing different VC systems in terms of their artifacts.

Although the tested countermeasure utilizing CQCC front-end and GMM backend was found to be sensitive to the choices of model parameters and acoustic features, our results indicate clear potential of spoofing countermeasure scores as a convenient and complementary tool to automatically assess the amount of audible and non-audible speech artifacts.

The obtained results are in a reasonable agreement with types of waveform generation methods used for VC systems. Our results indicate that the waveform filtering, SuperVP and

Griffin-Lim methods have relatively less artifacts. Since they are not included in the current ASVspoof datasets, we claim that it is also important to create new spoofing materials based on the methods and to train more robust anti-spoofing countermeasures in practice.

We also revealed that perceptually convincing VC samples based on Wavenet [25] in the VCC'18 have detectable artifacts. This implies that the current best VC samples may fool human ears but not necessarily the CM systems. There is no system that would perfectly fool both the humans and CM systems yet.

While our study serves as a proof-of-concept, we foresee several possible future directions. Firstly, it would be interesting to compare the performance to standard objective artifact measures used in assessing speech codecs and speech enhancement methods, and to other spoofing countermeasure front-ends besides CQCCs. Some alternative features may provide a stronger correlation to the MOS scores. Second, given the obvious issue related to selection and enriching of the training data with the latest VC techniques it might be relevant to consider one-class approaches that require human training speech only. Even if such approaches have had only moderate success in anti-spoofing [26] it would be interesting to revisit them in the context of artifact assessment.

7. Acknowledgements

We are grateful to iFlytek Ltd. for sponsoring the evaluation of the VCC 2018. This work was partially supported by MEXT KAKENHI Grant Numbers (15H01686, 16H06302, 17H04687, 17H06101), and Academy of Finland (proj. no. 309629).

8. References

[1] Seyed Hamidreza Mohammadi and Alexander Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] Yannis Stylianou, "Voice transformation: A survey," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, 2009, pp. 3585–3588.

[3] Tomi Kinnunen, Zhizheng Wu, Kong-Aik Lee, Filip Sedlak, Engsiong Chng, and Haizhou Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, 2012, pp. 4401–4404.

[4] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Voice conversion using input-to-output highway networks," *Information and Systems, IEICE Transactions on*, vol. E100.D, no. 8, pp. 1925–1928, 2017.

[6] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with wavenet-based waveform generation," in *Proc. Interspeech 2017*, 2017, pp. 1138–1142.

[7] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The voice conversion challenge 2016," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 1632–1636.

[8] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Accepted to Odyssey 2018*, 2018.

[9] Zhizheng Wu, Nicholas W. D. Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[10] Bryan L. Pellom and John H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, Arizona, USA, March 15-19, 1999*, 1999, pp. 837–840.

[11] Zhizheng Wu, Junichi Yamagishi, Tomi Kinnunen, Cemal Hanilçi, Md. Sahidullah, Aleksandr Sizov, Nicholas W. D. Evans, and Massimiliano Todisco, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *J. Sel. Topics Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[12] Pavel Korshunov, Sébastien Marcel, Hannah Muckenhirn, Andre R. Goncalves, A. G. Souza Mello, Ricardo P. Velloso Violato, Flávio O. Simões, M. U. Neto, Marcus de Assis Angeloni, José Augusto Stuchi, Heinrich Dinkel, Nanxin Chen, Yanmin Qian, Dipjyoti Paul, Goutam Saha, and Md. Sahidullah, "Overview of BTAS 2016 speaker anti-spoofing competition," in *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, 2016, pp. 1–6.

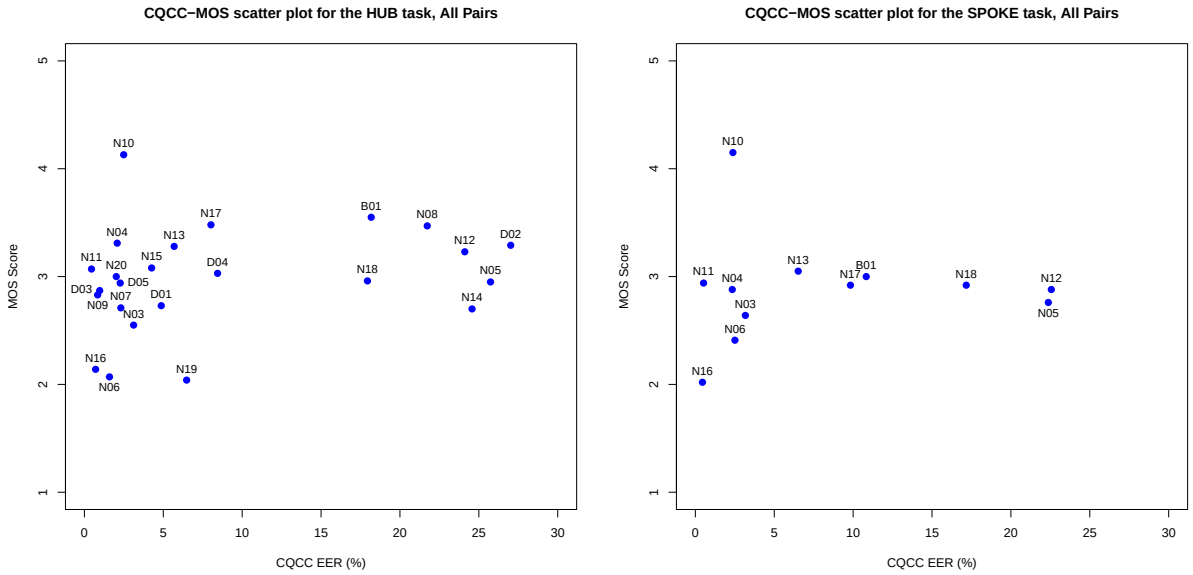
[13] Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi, "A comparison of features for synthetic speech detection," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2087–2091.

[14] Massimiliano Todisco, Héctor Delgado, and Nicholas Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey 2016: The Speaker and Language Recognition Workshop*, pp. 283–290.

[15] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Haizhou Li, "Front-end for antispoofing countermeasures in speaker verification: Scattering spectral decomposition," *J. Sel. Topics Signal Processing*, vol. 11, no. 4, pp. 632–643, 2017.

[16] Tanvina B. Patel and Hemant A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, 2015, pp. 2062–2066.

[17] Galina Lavrentyeva, Sergey Novoselov, Egor Malykh, Alexander Kozlov, Oleg Kudashev, and Vadim



(a) (b)

Figure 2: Scatter plot of objective vs. subjective quality of the VCC'18 challenge entries for **(a)** the HUB task and **(b)** the SPOKE task. The vertical axis represents mean opinion score (MOS) of subjective quality ratings of the samples of each challenge entry, while the horizontal axis represents equal error rate (EER, %) of spoofing countermeasure optimized to detect human and artificial speech. Therefore, the *higher* the EER value, the more confused the countermeasure is in telling apart the converted samples from authentic human speech, implying higher quality of the samples. The 'ideal' values would be MOS = 5.0 and EER = 50 %.

Shchemelinin, "Audio replay attack detection with deep learning frameworks," in *Proc. Interspeech 2017*, 2017, pp. 82–86.

[18] David A. van Leeuwen and Niko Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 1619–1623.

[19] Judith Brown, "Calculation of a constant q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[20] Seyed Omid Sadjadi, Malcolm Slaney, and Larry Heck, "MSR identity toolbox: A MATLAB toolbox for speaker recognition research (v 1.0)," https://www.microsoft.com/en-us/download/confirmation.aspx?id=52279, 2013.

[21] Pavel Korshunov and Sébastien Marcel, "Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations," *J. Sel. Topics Signal Processing*, vol. 11, no. 4, pp. 695–705, 2017.

[22] Cemal Hanilçi, Tomi Kinnunen, Md. Sahidullah, and Aleksandr Sizov, "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83–97, 2016.

[23] Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Md. Sahidullah, Wei Ming Liu, Federico Alegre, Tomi Kinnunen, and Benoit G. B. Fauve, "Impact of bandwidth and channel variation on presentation attack detection for speaker verification," in *International Conference of the Biometrics Special Interest Group, BIOSIG 2017, Darmstadt, Germany, September 20-22, 2017*, 2017, pp. 1–6.

[24] Gautham J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? - A dataset, insights, and challenges," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1006–1010, 2015.

[25] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv pre-print*, 2016.

[26] Federico Alegre, Asmaa Amehraye, and Nicholas W. D. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*, 2013, pp. 1–8.