Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions

Rohan Kumar Das¹, Tomi Kinnunen², Wen-Chin Huang³, Zhenhua Ling⁴, Junichi Yamagishi⁵, Yi Zhao⁵, Xiaohai Tian¹, Tomoki Toda³

¹National University of Singapore, Singapore ²University of Eastern Finland, Finland ³Nagoya University, Japan ⁴University of Science and Technology of China, China ⁵National Institute of Informatics, Japan

vcc2020@vc-challenge.org

Abstract

The Voice Conversion Challenge 2020 is the third edition under its flagship that promotes intra-lingual semiparallel and cross-lingual voice conversion (VC). While the primary evaluation of the challenge submissions was done through crowd-sourced listening tests, we also performed an objective assessment of the submitted systems. The aim of the objective assessment is to provide complementary performance analysis that may be more beneficial than the time-consuming listening tests. In this study, we examined five types of objective assessments using automatic speaker verification (ASV), neural speaker embeddings, spoofing countermeasures, predicted mean opinion scores (MOS), and automatic speech recognition (ASR). Each of these objective measures assesses the VC output along different aspects. We observed that the correlations of these objective assessments with the subjective results were high for ASV, neural speaker embedding, and ASR, which makes them more influential for predicting subjective test results. In addition, we performed spoofing assessments on the submitted systems and identified some of the VC methods showing a potentially high security risk.

Index Terms: Voice Conversion Challenge 2020, objective evaluation, subjective rating prediction, spoofing assessment

1. Introduction

Voice conversion (VC), which refers to the digital cloning of a person’s voice, can be used to modify an audio waveform so that it appears as if spoken by someone else (target) than the original speaker (source). VC is useful in many applications such as customizing audio book and avatar voices, dubbing, the movie industry, teleconferencing, singing voice modification, voice restoration after surgery, and the cloning of voices of historical persons. Since VC technology involves identity conversion, it can also be used to protect the privacy of individuals on social media and in sensitive interviews. For the same reason, VC also enables spoofing (fooling) voice biometric systems and therefore has potential security implications.

VCC 2020 is the 3rd edition of the Voice Conversion Challenge (VCC). While the general background and subjective results are provided in [1], this study focuses on complementary objective evaluation results.

Conventionally, the target of VC technology is human listeners, so subjective assessment has been the primary method of assessment in all the VCC challenges. On the other hand, progress has been made recently in research fields relevant to objective evaluation assessment, and human perception prediction and spoofing performance assessment against automatic speaker verification (ASV) systems are increasingly being used. The former is utilized to model and predict human perceptions automatically. Until recently, predicting human perception on synthetic speech has been challenging, but new research [2–9] has demonstrated that the advanced deep learning models and large amounts of paired data of synthetic speech and associated human judgement scores can lead to data-driven objective models that can predict human perception on synthetic speech to a certain extent. There are also several studies on using automatic speech recognition (ASR) as a proxy for subjective intelligibility estimation [10].

The latter pertains to spoofing and anti-spoofing research for ASV. One goal of VC is to produce convincing mimicry of specific target speaker voices, and it is widely known that VC can fool (spooﬁng) unprotected ASV systems [11, 12]. Therefore, for increasing security and robustness, ASV systems normally adopt spoofing countermeasures, which are designed to learn the distinguishing artifacts present in spoofed audio produced by VC from human speech. We assume that spoofing performance against ASV may be correlated with the speaker similarity of VC systems and that the countermeasure (CM) performance represents the amount of artifacts produced by VC systems, which may or may not be audible to humans. Note that these systems are designed and optimized for discrimination by machines, and as such, the performances of ASV and CM may be different from human perceptions.

With the above as our motivation, we provide an array of complementary objective results representative of recent objective evaluation techniques. Our tools include

- text-independent ASV [13] for speaker similarity,
- text-independent CM [11] for real-vs.-fake assessment,
- automatic MOS prediction [6] for quality, and
- ASR for intelligibility.

To the best of our knowledge, these four metrics have never before been examined as a group within any single VC study. Using these metrics, this paper investigates the two questions below:

- Can the metrics predict human judgements on naturalness and speaker similarity?
- Which VC technology has the highest spoofing risk for ASV and CM?

In Section 2 of this paper, we give an overview of the motivation behind each of the objective evaluation metrics. Their implementation details are described in Section 3. Correlations with human judgements are analyzed in Section 4, and spoofing performance against ASV and CM is discussed in Section...
5. We conclude in Section 6 with a brief summary and mention of future work.

2. Methodology

Objective metrics for speech signals can be categorized into intrusive and non-intrusive assessment methods. The former uses a ground-truth natural clean audio that has the same linguistic content as an input speech as a reference. The latter does not use any reference. A summary of the objective metrics used in this paper is provided in Table 1. They are all non-intrusive metrics. The following subsections provide the motivation and general description of each method.

2.1. Speaker Similarity: Automatic Speaker Verification

We assess speaker similarity using ASV. An ASV system compares a test utterance of an unknown speaker with a hypothesized speaker’s training utterance(s) and then outputs a speaker similarity score $s \in \mathbb{R}$. Higher scores indicate support for the same speaker hypothesis and lower scores for the different speakers. A hard decision is obtained by comparing $s$ with a pre-set verification threshold, $\tau^{asv}$. The speakers are declared to be the same if $s > \tau^{asv}$ (otherwise, different).

When VC methods achieve good mimicry of specific target speaker voices, the above score becomes higher and hence the converted audio will be judged as the same speaker. Therefore, we use the impact of VC on the ASV error rates as speaker similarity metrics here. For each VC system, we report three different kinds of ASV error:

1. Equal error rate (EER): error rate at $\tau^{asv}$ at which the spoof false acceptance rate and target speaker miss rate equal each other;
2. False acceptance rate of target ($P_{fa}^{tar}$): proportion of converted utterances declared as the targeted speaker;
3. Miss rate of source ($P_{miss}^{src}$): proportion of converted utterances not declared as the original source speaker.

The first one gauges the ASV system’s general accuracy in differentiating converted source utterances from real target speaker utterances (alternatively, the effectiveness of a VC in fooling ASV). The second metric also gauges the VC system’s ability to fool the ASV system, with the difference that our false acceptance rate computation uses a threshold fixed prior to observing any VC samples. The same holds for the third metric, source miss rate, which gauges the VC system’s ability to de-identify the source speaker, in terms of ASV (alternatively, the ASV system’s inability to re-identify the source speaker). The reference values for ideal VC and useless VC are $(EER, P_{fa}^{tar}, P_{miss}^{src}) = (\frac{1}{2}, 1, 1)$ and $(EER, P_{fa}^{tar}, P_{miss}^{src}) = (0, 0, 0)$, respectively.

The second and third error rates defined above are obtained by fixing $\tau^{asv}$ before observing the converted utterances. The ASV system is optimized to differentiate real human same-speaker and different-speaker trials, but no VC samples are used in optimizing it. Thus, $\tau^{asv}$ is fixed using non-converted (natural) data only and remains fixed for the VC test. In practice, we set $\tau^{asv}$ at the EER operating point at which the (natural speech) miss and false alarm rates equal each other.

2.2. Speaker Similarity: Neural Speaker Embedding

Strictly speaking, the above ASV scoring procedures are different from how our subjects evaluated speaker similarity in the main listening test [1]. In the listening test, they were asked to listen to a reference audio (which had different linguistic content from converted audio) and judge the speaker similarity of the converted audio to one of the reference audio files.

Therefore, we computed the cosine similarity of speaker embedding vectors as well as the ASV scores. We extracted speaker embedding vectors from both converted audio and one of the reference audio files using the same tool as the above ASV system and measured their cosine similarity, $\cos_{sim}(A, B) = A \cdot B / \|A\| \|B\|$, where $A$ and $B$ are the speaker embedding vectors obtained from the converted audio and reference audio, respectively. Note that this is still technically a non-intrusive assessment because the reference audio for this measurement is different from the ground-truth natural clean audio that has the same linguistic content as an input speech.

2.3. Artifact Assessment: Spoofing Countermeasures

The spoofing countermeasures play an imperative role in defending against various attacks to ASV systems. In general, such countermeasures are designed to learn the distinguishing artifacts present in the natural human speech against different kinds of generated/spoofed speech to identify the spoofing attacks. Therefore, the artifact assessment using a spoofing countermeasure for the resultant speech of various VC systems could indicate the amount of artifacts of the converted speech.

The performance of spoofing countermeasures is normally evaluated in terms of EER, similar to that of the ASV systems. However, unlike ASV, any human speech is considered as target trials and the converted/spoofed speech serve as non-

---

Table 1: Summary of evaluation metrics used in assessing VCC 2020 submissions. ASV: automatic speaker verification, EER: equal error rate, MOS: mean opinion score, LCNN: light convolutional neural network, ASR: automatic speech recognition, WER: word error rate, CM: countermeasure.

| Metric       | Type of measure               | Measurement tool   | Implementation                              | Metric interpretation                  |
|--------------|-------------------------------|--------------------|---------------------------------------------|----------------------------------------|
| ASV EER      | Conv. src ↔ tgt similarity    | ASV                | Kaldi x-vector [13]                         | Not similar = 0% ... 50% = Similar     |
| $P_{fa}^{tar}$ | Conv. src ↔ tgt similarity   | ASV                | Kaldi x-vector [13]                         | Not similar = 0% ... 100% = Similar    |
| $P_{miss}^{src}$ | Conv. src ↔ src similarity  | ASV                | Kaldi x-vector [13]                         | Similar = 0% ... 100% = Not similar   |
| Cosine      | Conv. src ↔ tgt similarity    | Speaker embedding  | Kaldi x-vector [13]                         | Not similar = -1 ... 1 = Similar       |
| CM EER      | Artifact assessment           | Spoofer CM         | LCNN [14]                                   | Fake = 0% ... 50% = Real               |
| MOSNet      | Quality                       | Objective MOS      | MOSNet [6]                                  | Lowest = 1 ... 5 = Highest             |
| ASR WER     | Intelligibility               | ASR                | Seq2seq with attention [15]                 | Perfect = 0%...100% = Unintelligible  |

1Theoretically, EER is constrained between 0 and $\frac{1}{2}$ for any binary classification task, including ASV. Values larger than $\frac{1}{2}$ indicate decisions worse than random guessing (label flip). In practice, the values may exceed $\frac{1}{2}$ due to data anomalies.
target trials. In the context of VC outputs, a high EER indicates the generation of more human-like speech, whereas a low EER indicates that the converted speech is inclined towards the characteristics of artificially generated speech.

2.4. Quality: Objective Mean Opinion Scores

The mean opinion score (MOS) used for subjective quality assessment is a numerical measure of the human-judged overall quality of audio, typically in the range of 1.5, where 1 is the lowest perceived quality and 5 is the highest. The objective MOS is used to assess speech quality by predicting and approximating the human assessment from an input audio. This technique has a long history and several metrics have been proposed for speech coding and telephony. Famous metrics include PESQ [16] and POLQA [17], both of which are intrusive assessments. There is also p536 [18], a metric for non-intrusive assessment, but it is not designed for evaluating the quality of synthetic speech or converted speech.

When a large amount of paired data of synthetic speech and associated human judgement scores is available, we can view the objective MOS as a machine learning-based regression problem. Various deep learning models to predict the MOS values utilizing listening test data collected from the Blizzard Challenge [19], our previous VCCs [20, 21], or the ASVspoof Challenge [22] for supervision have been proposed. More specifically, [2] conducted prediction of the Blizzard Challenge’s results, [6, 8, 9] reported predictions of VCC’s results, and [7] reported prediction results of ASVspoof’s listening test results. While all of them exhibited moderate correlations with human judgments, it is still unknown whether these models can be generalized to new speech synthesis methods that are not already included in the databases.

Among these models, we have chosen to examine MOSNet [6], which is a deep learning-based non-intrusive assessor, since it is reported that its prediction has a moderate correlation with the subjective evaluation done for the VCC 2018.

2.5. Intelligibility: Automatic Speech Recognition

Although some of the recent VC methods can achieve high naturalness and similarity of converted speech, they may degrade the intelligibility of converted speech. For example, in the recognition-synthesis approach [23-26] to VC, an ASR model is usually adopted to extract linguistic-related features, e.g., phonetic posteriorgrams (PPGs) [24] or bottleneck features [26], from source speech. In this case, recognition errors are inevitable, which may degrade the intelligibility of the converted speech. Moreover, in the VC methods using sequence-to-sequence acoustic models [27], the failed alignment may lead to repetition and deletion of speech segments, especially when the amount of training data is limited [28].

Considering the cost of conducting subjective intelligibility evaluations for all conversion pairs, we adopted the word error rate (WER) of ASR as an objective metric on the intelligibility evaluations for all conversion pairs, we adopted the word error rate (WER) of ASR as an objective metric on the intelligibility of converted speech. The training x-vectors of each utterance are used in [6]. Specifically, the raw magnitude spectrogram was first extracted from the converted speech and used as the input feature. The main model consisted of 12 convolution layers, one bidirectional long-short term memory layer, and two fully connected layers followed by a global averaging layer that pooled the frame-level scores to generate the final utterance level predicted MOS. The whole network was trained to regress the listening test scores by minimizing the mean square loss.

For training the MOSNet, we used two datasets: one composed of listening test data collected for the VCC 2018 [21] and the other of listening test data collected for the ASVspoof 2019 [22]. The former contains many of the VC systems available in 2018 and the latter contains more recent speech synthesis and VC methods available from 2019. They are referred to as MOSNet (vcc18)² and MOSNet (asvspoof19)³, respectively.

3. Implementation Details

3.1. ASV and Cosine Distance of Speaker Embeddings

Our ASV system utilizes x-vector [29]-based deep speaker embeddings. We use Kaldi’s [30] recipe [13] trained on VoxCeleb data [31]. The system uses a time-delay neural network model (TDNN) trained with cross-entropy loss (treating training speakers as classes) to extract one 512-dimensional deep speaker embedding per utterance. The speaker similarity score is computed using probabilistic linear discriminant analysis (PLDA).

The system is used as a scoring tool without specific modifications (e.g., domain adaptation) for the VCC 2020 data. The source and target speaker reference models are obtained from the respective training utterances provided to the challenge participants. The training x-vectors of each utterance are used to form one averaged x-vector per speaker. Test utterance x-vectors are then scored against these averaged models. For pre-VC ASV tests (required when setting the ASV threshold), we use the original source test data (provided to challenge participants) and the target speaker reference data (not provided to challenge participants). For the VC tests, we replace the original source utterances with their VC-processed versions.

For calculating the cosine distance of the speaker embeddings, we use the same Kaldi-based x-vector extractor as the ASV system. The x-vector dimensions were reduced to 200 using LDA before we compute the cosine similarity between converted speech and natural speech. We then calculate the averaged value per system.

3.2. Spoofing Countermeasure

We used the light convolutional neural network (LCNN)-based system as a spoofing countermeasure in our studies [14]. The system considers 60-dimensional (20-static+20-Δ+20-ΔΔ) linear frequency cepstral coefficient (LFCC) features as the input [32]. The training set of the ASVspoof 2019 logical access corpus is used to build the model [22]. The detailed architecture and implementation of the LCNN system is available in [33]. We considered the utterances from the training set of the VCC 2020 database as the bona fide trials, whereas the submissions from various teams constitute the spoof trials for evaluating the performance of every system submitted to the challenge.

3.3. MOSNet

Our MOSNet model architecture followed the original setting in [6]. Specifically, the raw magnitude spectrogram was first extracted from the converted speech and used as the input feature. The main model consisted of 12 convolution layers, one bidirectional long-short term memory layer, and two fully connected layers followed by a global averaging layer that pooled the frame-level scores to generate the final utterance level predicted MOS. The whole network was trained to regress the listening test scores by minimizing the mean square loss.

For training the MOSNet, we used two datasets: one composed of listening test data collected for the VCC 2018 [21] and the other of listening test data collected for the ASVspoof 2019 [22]. The former contains many of the VC systems available in 2018 and the latter contains more recent speech synthesis and VC methods available from 2019. They are referred to as MOSNet (vcc18)² and MOSNet (asvspoof19)³, respectively.

3.4. ASR Engine

The ASR engine was a prototype system developed by iFlytek. It features a state-of-the art end-to-end neural network-based

²https://github.com/lochenchou/MOSNet
³https://github.com/rhoposit/MOS_Estimation2
ASR architecture and was trained using 10,000 hours of recordings and GB-level texts for language modeling. The vocabulary size was around 200,000. WERs were calculated by the HResults tool in HTK using manual transcriptions as ground truth and considering substitutions, deletions, and insertions.

4. Objective Evaluation Results for Each Submitted VC System

We used converted audio produced by each of the submitted VC systems for Tasks 1 and 2 and assessed the objective evaluation metrics described earlier. It is noted that the Tasks 1 and 2 are intra-lingual semi-parallel and cross-lingual VC tasks, respectively. The results for each VC system for Tasks 1 and 2 are summarized in Tables 2 and 3, respectively.

ASV: The ASV results are shown in the first to fourth columns in the tables. Concerning false acceptance rates on Task 1, 20 (out of 31) systems achieved error rates higher than 90%. Further, the top-5 systems obtained perfect results (100% false acceptance). Many other teams obtained near-perfect results. Concerning the miss rate of converted source speakers, it was higher than 90% for all teams except three (06, 08, 12). To sum up, the top systems all achieved similar results, and the majority of the systems achieved a high target speaker similarity. Nearly all systems managed to successfully move the converted voice ‘away’ from the original source. The results were more varied for Task 2, however: only ten systems yielded false acceptance rates above 90%, and there was substantial variation across the systems. The miss rates were generally worse than those in Task 1 for most systems, although, similar to Task 1, for most systems they were reasonably high.

Spoofing Countermeasures: The fifth column of the tables shows the performance of the LCNN-based spoofing countermeasure for all the submitted systems on both tasks of VCC 2020. We can see that most of the teams achieved a high EER, which indicates that the VC systems were able to generate natural human-like speech that was not easily detectable by the spoofing countermeasure. In addition, the performance trends of the spoofing countermeasure for most teams were similar in both tasks, excluding four teams (08, 18, 22 and 23). They also showed a relatively higher EER for Task 2 than Task 1, which might be a result of the very different settings used for both tasks by those teams.

MOSNet: The MOSNet predictions of all systems are shown in the sixth column in the tables. We can see that the MOSNet predictions fell between 2.5 to 4.5, while the ground truth MOS typically ranged from 1.0 to 4.5, indicating that the overall variance of the MOSNet predictions was rather small. We can also see that the scores of each team for Tasks 1 and 2 were similar. This is consistent with the fact that most teams utilized the same system for both tasks.

WER: The ASR WERs of all systems are shown in the final column in the tables. We can observe a large variance of WERs among teams in both tasks. For example, in Task 1, seven teams had WERs that were lower than 5%, while seven other teams had WERs that were higher than 50%. This indicates the diversity of the conversion methods adopted by different teams. After subjectively examining a few samples from the teams with high WERs, we can clearly perceive their intelligibility degradation. Comparing the WERs in Tasks 1 and 2, as expected, most teams achieved a higher WER on Task 2, which indicates the difficulty of cross-lingual VC.

5. Can the Metrics Predict Human Judgements on VC Speech?

5.1. Correlation with subjective evaluation results

In this section, we investigate our first question, “Can the metrics predict human judgements on the naturalness and speaker similarity of converted audio submitted for VCC 2020?”

In the VCC 2020, we conducted two large-scale crowdsourced listening tests on the naturalness and speaker similarity of converted speech. The first test was done by 68 native English listeners (32 female, 33 male, and 3 unknown) and the second by 206 native Japanese listeners (96 male and 110 female). More details are described in [1]. Using these listening test results, we measured the correlations of each objective evaluation metric with the subjective evaluation results by the English and Japanese listeners.

More specifically, we created scatter plots matching each of the objective metrics at the system level and each of the subjective evaluation results, then calculated the Pearson correlation coefficients. The scatter plots are shown in Appendix B.

Table 4 shows the Pearson correlation coefficients with subjective evaluation results for each metric along with their p-values. The top and bottom tables show the correlations with the subjective scores obtained from the English and Japanese listeners, respectively. We first summarize the correlation analysis using the English listeners and then discuss the difference between this case and the Japanese one.

Subjective quality rating: From Table 4, we can see that the ASV-related metrics (EER, Pfa), MOSNet (vcc18, asvspoof19), and ASR WER had moderately positive or negative correlations with the subjective quality ratings in Task 1, while cosine distance, MOSNet (asvspoof19), and ASR WER had moderate correlations in Task 2. These findings are statistically significant. While it is slightly surprising to see the ASV-related metrics and cosine distance were correlated with the subjective quality ratings, we assume this stems from the fact that human judgements on quality and speaker similarity are not independent. The fact that the ASV-related metrics had higher correlations with the speaker similarity ratings also underpins this. We can also see that MOSNet (asvspoof19) had higher correlations than MOSNet (vcc18). This was because the asvspoof19 dataset contains more diverse and new speech generation methods than the VCC18 dataset, which demonstrates the importance of choosing the appropriate training dataset for MOSNet.

Subjective speaker similarity rating: We can see that all of the ASV-related metrics (EER, Pfa, cosine distance) had strong correlations with subjective speaker similarity ratings in both tasks. Among these, the EER had the highest correlations ($r = 0.89$ for Task 1 and $r = 0.90$ for Task 2). MOSNet (vcc18, asvspoof19) had a moderate correlation with the speaker similarity ratings in Task 1, but its correlation in Task 2 was not statistically significant.

English listeners vs. Japanese listeners: Next, we analyzed the differences between the English listeners’ case and the Japanese one. We can see that the general tendencies were the same in both cases; that is, ASV-related metrics had strong correlations with subjective speaker similarity ratings, and MOSNet and ASR WER had moderately negative correlations with subjective quality ratings. Two minor differences were that the cosine distance had a slightly higher correlation than ASV EER for Task 2 and that MOSNet (asvspoof19) had a higher correlation than ASR WER for Task 2. These differences were marginal.
Table 2: Performance of objective measures for Task 1 (intra-lingual semi-parallel VC). Red cells indicate top-5 systems (including ties) for each metric.

| Team ID | ASV EER (%) | ASR Pfa (%) | ASR Pmiss (%) | Cosine | CM EER (%) | MOSNet (vcc18) | MOSNet (asvspoof19) | ASR WER (%) |
|---------|-------------|-------------|---------------|--------|------------|----------------|---------------------|-------------|
| T01     | 24.00       | 98.25       | 100.00        | 0.95   | 22.47      | 4.57           | 3.55                | 22.78       |
| T02     | 14.00       | 87.50       | 100.00        | 0.86   | 26.74      | 3.57           | 3.22                | 12.32       |
| T03     | 23.00       | 82.00       | 99.75         | 0.90   | 0.78       | 3.37           | 3.64                | 80.26       |
| T04     | 45.13       | 99.00       | 100.00        | 0.97   | 38.30      | 3.92           | 3.25                | 22.84       |
| T05     | 0.00        | 0.00        | 21.75         | 0.72   | 14.77      | 2.65           | 2.99                | 3.65        |
| T07     | 48.50       | 99.75       | 100.00        | 0.96   | 43.48      | 3.73           | 3.61                | 18.08       |
| T08     | 0.50        | 0.50        | 78.25         | 0.76   | 37.97      | 2.89           | 2.86                | 6.95        |
| T09     | 19.00       | 86.25       | 100.00        | 0.91   | 7.97       | 3.71           | 3.17                | 62.76       |
| T10     | 51.00       | 100.00      | 100.00        | 0.98   | 43.98      | 3.90           | 3.70                | 4.12        |
| T11     | 38.50       | 99.00       | 99.50         | 0.94   | 42.75      | 4.27           | 4.17                | 5.43        |
| T12     | 0.00        | 0.00        | 9.00          | 0.45   | 31.46      | 3.02           | 3.10                | 38.55       |
| T13     | 37.00       | 97.25       | 99.75         | 0.94   | 28.25      | 3.44           | 3.30                | 9.70        |
| T14     | 1.00        | 6.00        | 99.50         | 0.76   | 65.96      | 2.85           | 2.47                | 19.77       |
| T16     | 33.00       | 97.00       | 100.00        | 0.93   | 36.51      | 3.33           | 3.33                | 21.52       |
| T17     | 7.50        | 34.25       | 100.00        | 0.80   | 47.24      | 2.95           | 3.11                | 52.07       |
| T18     | 14.00       | 72.50       | 100.00        | 0.91   | 25.50      | 2.65           | 2.61                | 55.58       |
| T19     | 33.63       | 98.50       | 100.00        | 0.94   | 42.75      | 3.08           | 3.39                | 65.80       |
| T20     | 24.00       | 95.00       | 100.00        | 0.94   | 32.24      | 3.75           | 3.37                | 22.58       |
| T21     | 2.00        | 13.75       | 98.50         | 0.75   | 47.82      | 3.86           | 3.45                | 30.84       |
| T22     | 52.00       | 100.00      | 100.00        | 0.96   | 33.76      | 3.63           | 3.69                | 6.69        |
| T23     | 45.00       | 99.75       | 100.00        | 0.94   | 32.02      | 3.51           | 3.72                | 19.28       |
| T24     | 25.50       | 98.50       | 100.00        | 0.91   | 20.50      | 3.54           | 3.52                | 23.83       |
| T25     | 33.50       | 98.50       | 98.75         | 0.96   | 27.52      | 3.58           | 3.67                | 4.82        |
| T26     | 3.63        | 53.00       | 99.25         | 0.86   | 18.98      | 3.76           | 3.38                | 28.04       |
| T27     | 37.50       | 100.00      | 100.00        | 0.95   | 19.49      | 3.41           | 3.31                | 34.82       |
| T28     | 34.50       | 96.00       | 99.75         | 0.95   | 32.70      | 3.58           | 3.25                | 96.17       |
| T29     | 45.50       | 100.00      | 100.00        | 0.96   | 34.44      | 3.94           | 3.71                | 8.47        |
| T30     | 46.00       | 99.75       | 100.00        | 0.97   | 2.02       | 3.72           | 3.61                | 77.80       |
| T31     | 31.63       | 99.50       | 100.00        | 0.92   | 25.45      | 3.27           | 2.66                | 77.80       |
| T32     | 18.00       | 95.00       | 100.00        | 0.94   | 30.55      | 3.48           | 3.55                | 4.21        |
| T33     | 43.13       | 100.00      | 100.00        | 0.96   | 33.25      | 3.55           | 3.72                | 9.64        |

5.2. Analysis of objective speaker similarity scores of top-ranked VC submissions

We have demonstrated that the ASV metrics, in particular EER, had a strong correlation with subjective ratings on speaker similarity. Therefore, it should prove fruitful to analyze the top-ranked VC submissions further, as their subjective speaker similarity in Task 1 was as good as the target speakers according to the listening test results [1] and hence we cannot obtain meaningful differences between the top submissions from the listening test only.

As reported in [1], eight VC systems (T10, T22, T27, T13, T33, T29, T29, and T07) had statistically fewer significant differences from human speech. As expected, some of these differences related to the EERs (as shown in Table 2), although all of them had very high EERs above 35%. In particular, T10 and T22 had slightly higher EERs than T30, the chance speaker, and hence it is expected that T10 and T22 have slightly ‘emphasized’ speaker characteristics compared to real target speakers. T23, T29, and T07 also had EERs between 45% and 48% and were higher than T27, T13, and T33.
5.3. Analysis of prediction results for each language spoken by target speakers

Task 2 of VCC 2020 is a cross-lingual VC task and the target speaker’s speech contains utterances in German, Finnish, and Mandarin. As reported in [1], this factor affected the performance of the VC systems built by challenge participants. The converted audio files for German target speakers had the highest naturalness and speaker similarity, while those for Mandarin had the lowest. Therefore, we want to confirm whether the objective metrics can capture such score variations and predict the listening test scores for each language.

A breakdown of the Pearson correlation coefficients per language is provided in Table 5. The top and bottom tables show correlations with the subjective scores from the English and Japanese listeners, respectively. The results for individual submissions are shown in Appendix C.

As we can see, the correlations of the ASV metrics and ASR WER were similar and stable across the three languages. Again, MOS ratings were correlated with ASR WER and speaker similarity ratings were well correlated with ASV metrics. We can also see that the MOS ratings done by Japanese listeners had weaker correlations with ASR WER than those done by English listeners.

5.4. Can we predict subjective evaluation results better by using combinations of the objective metrics?

Next, we carried out multiple linear regression analysis to determine whether the prediction accuracy on subjective scores could be improved by combining several objective metrics. Since the ASV metrics reported in previous subsections have “multicollinearity”, we selected ASV EER only for this analysis. Also, we chose MOSNet (asvspoof19) over MOSnet (vcc18) since the former had higher correlation with the subjective scores (as shown in Table 4). The estimated coefficients and the statistics are listed in Table 6, where the top and bottom parts show subjective scores from English and Japanese listeners, respectively.

From the table, by comparing the adjusted R-squared values and the Pearson correlation coefficients of Table 4, we can see that the prediction accuracy on subjective quality rating scores can be improved by combining multiple objective metrics for all of Task 1 and for the quality rating of Task 2. The significant explainable variables for MOS were ASV EER and ASR WER for Task 1 and MOSNet (asvspoof19) and ASR WER for Task 2. This was the case for both the English and Japanese listeners’ scores. However, this was not the case for the subjective speaker similarity rating, and only the coefficient for ASV EER was statistically significant. This is presumably because the ASV EER itself had sufficiently high correlations.

---

Table 4: Pearson correlation coefficients with subjective evaluation results for each metric and (p-values). Top: correlation with subjective scores from English listeners. Bottom: correlation with subjective scores from Japanese listeners. Bold font indicates the highest correlation among the objective metrics.

| Subjective score (ENG) | ASV EER (%) | ASV PFA (%) | Cosine distance | Countermeasure EER (%) | MOSNet (vcc18) | MOSNet (asvspoof19) | ASR WER (%) |
|------------------------|-------------|-------------|-----------------|------------------------|----------------|---------------------|-------------|
| Task 1 MOS             | 0.70 (p < 0.01) | 0.53 (p < 0.01) | 0.42 (p < 0.01) | 0.00 (p > 0.01) | 0.52 (p < 0.01) | 0.66 (p < 0.01) | -0.65 (p < 0.01) |
| Task 1 SIM             | 0.89 (p < 0.01) | 0.82 (p < 0.01) | 0.85 (p < 0.01) | 0.07 (p > 0.01) | 0.54 (p < 0.01) | 0.61 (p < 0.01) | -0.18 (p > 0.01) |
| Task 2 MOS             | 0.34 (p > 0.01) | 0.26 (p < 0.01) | 0.59 (p < 0.01) | 0.27 (p > 0.01) | 0.43 (p > 0.01) | 0.58 (p < 0.01) | -0.73 (p < 0.01) |
| Task 2 SIM             | 0.90 (p < 0.01) | 0.86 (p < 0.01) | 0.82 (p < 0.01) | 0.19 (p > 0.01) | 0.23 (p > 0.01) | 0.32 (p > 0.01) | -0.14 (p > 0.01) |

Table 5: Breakdown per language of Pearson correlation coefficients with subjective evaluation results for each metric along with p-values. Top: correlation with subjective scores from English listeners. Bottom: correlation with subjective scores from Japanese listeners. Bold font indicates the highest correlation among the objective metrics. F, G, and M represent Finnish, German, and Mandarin target speakers, respectively.

| Subjective score (JPN) | ASV EER (%) | ASV PFA (%) | Cosine distance | Countermeasure EER (%) | MOSNet (vcc18) | MOSNet (asvspoof19) | ASR WER (%) |
|------------------------|-------------|-------------|-----------------|------------------------|----------------|---------------------|-------------|
| Task 2 MOS (F)         | 0.26 (p > 0.01) | 0.26 (p > 0.01) | 0.62 (p < 0.01) | 0.40 (p < 0.01) | 0.42 (p > 0.01) | 0.41 (p < 0.01) | -0.74 (p < 0.01) |
| Task 2 MOS (G)         | 0.35 (p < 0.01) | 0.27 (p > 0.01) | 0.45 (p > 0.01) | 0.16 (p > 0.01) | 0.47 (p > 0.01) | 0.47 (p > 0.01) | -0.71 (p < 0.01) |
| Task 2 MOS (M)         | 0.36 (p > 0.01) | 0.26 (p < 0.01) | 0.64 (p < 0.01) | 0.10 (p > 0.01) | 0.34 (p > 0.01) | 0.34 (p > 0.01) | -0.72 (p < 0.01) |
| Task 2 SIM (F)         | 0.08 (p < 0.01) | 0.18 (p < 0.01) | 0.70 (p > 0.01) | 0.27 (p > 0.01) | 0.25 (p > 0.01) | 0.26 (p > 0.01) | -0.09 (p < 0.01) |
| Task 2 SIM (G)         | 0.02 (p < 0.01) | 0.02 (p < 0.01) | 0.84 (p < 0.01) | 0.08 (p < 0.01) | 0.38 (p > 0.01) | 0.37 (p < 0.01) | -0.26 (p > 0.01) |
| Task 2 SIM (M)         | 0.08 (p < 0.01) | 0.08 (p < 0.01) | 0.74 (p < 0.01) | 0.09 (p > 0.01) | 0.06 (p < 0.01) | 0.06 (p > 0.01) | -0.02 (p > 0.01) |

| Subjective score (JPN) | ASV EER (%) | ASV PFA (%) | Cosine distance | Countermeasure EER (%) | MOSNet (vcc18) | MOSNet (asvspoof19) | ASR WER (%) |
|------------------------|-------------|-------------|-----------------|------------------------|----------------|---------------------|-------------|
| Task 2 MOS (F)         | 0.34 (p > 0.01) | 0.34 (p > 0.01) | 0.67 (p < 0.01) | 0.42 (p > 0.01) | 0.46 (p > 0.01) | 0.46 (p > 0.01) | -0.62 (p > 0.01) |
| Task 2 MOS (G)         | 0.41 (p > 0.01) | 0.32 (p > 0.01) | 0.51 (p < 0.01) | 0.20 (p > 0.01) | 0.50 (p < 0.01) | 0.50 (p < 0.01) | -0.55 (p > 0.01) |
| Task 2 MOS (M)         | 0.40 (p > 0.01) | 0.25 (p > 0.01) | 0.65 (p > 0.01) | 0.20 (p > 0.01) | 0.43 (p > 0.01) | 0.42 (p > 0.01) | -0.56 (p > 0.01) |
| Task 2 SIM (F)         | 0.75 (p < 0.01) | 0.75 (p < 0.01) | 0.81 (p < 0.01) | 0.36 (p > 0.01) | 0.27 (p > 0.01) | 0.27 (p > 0.01) | -0.29 (p > 0.01) |
| Task 2 SIM (G)         | 0.82 (p < 0.01) | 0.82 (p < 0.01) | 0.85 (p < 0.01) | 0.18 (p > 0.01) | 0.34 (p > 0.01) | 0.35 (p > 0.01) | -0.27 (p > 0.01) |
| Task 2 SIM (M)         | 0.81 (p < 0.01) | 0.77 (p < 0.01) | 0.78 (p < 0.01) | 0.25 (p > 0.01) | 0.11 (p > 0.01) | 0.11 (p > 0.01) | -0.19 (p > 0.01) |
finding is consistent with the correlation analysis discussed in the previous section.

We can also see that the MOS score regression models in Task 2 had lower adjusted R-squared values (0.74 and 0.68) compared to the other regression results. This clearly suggests that predicting the naturalness score in the cross-lingual setting is harder, and needs to be explained by more factors.

5.5. Discussion

The results of our analysis demonstrated that MOS ratings were correlated with ASR WER or MOSNet (asvspoof19) and speaker similarity ratings were well correlated with ASV metrics, and that they were complementary for quality rating.

Contrary to our expectations, only the ASV and ASR models trained using human speech data became important explainable variables for predicting subjective ratings, and the CM models using synthesized speech or converted speech did not. This might be due to the many new types of waveform generation methods adopted by the challenge participants. As reported in [1], ten types of vocoders were used for VCC 2020, many of which were not included in the training databases used for the CM model.

We should point out that the correlations reported in this paper are at the system level and as such represent neither the quality nor the speaker similarity of individual sentences. Investigating sentence-level score predictions will be the focus of future work.

6. Spoofing Performance Assessment

In this section, we address our second question, “Which VC technology has the highest spoofing risk for ASV and CM?” If we can answer this question and identify which VC methods the current ASV and CM systems are most vulnerable to, the speaker recognition community can use our results as a guideline to efficiently prepare additional training data to enhance the robustness of ASV and CM systems. Once spoofed data produced by the missing VC technologies is added to training databases for the CM systems and the VC technologies become “known” from the CM perspective, their discrimination normally becomes much easier. Therefore, we focus on the ASV and CM metrics in this section.

6.1. t-DCF-based spoofing performance analysis for each submitted VC system

The ASV and CM results in Tables 2 and 3 were reported in isolation from each other. For “spoofing performance” assessment, we consider the ASV and CM results jointly. Specifically, we envision a cascaded (tandem) system [34] where a CM system is placed before the ASV, with the aim of preventing spoofing attacks from reaching the ASV system. VC audio files may be rejected by a CM system if the audio contains detectable artifacts. Even if VC audio files are passed on to the CM system, they may still be rejected by the ASV if their speaker similarity is not close enough to the target speakers. Further, while our primary interest is the performance degradation due to falsely accepted VC spoofing attacks, the tandem system is also prone to two other types of errors. First, it may accept another similar-appearing human user (a non-target) as a target. Second, either CM or ASV may reject the actual target user. Using an overly sloppy CM or ASV system leads to compromised security, while conversely, an overly aggressive CM (or ASV) leads to reduced user convenience. To assess the joint performance of CM and ASV, we adopt a new metric called (minimum) tandem detection cost function (t-DCF) [34]. It combines all the ASV and CM results into a single cost value for each submitted VC system.

Unlike the parameter-free metrics considered above, t-DCF is a parameterized cost that makes the modeling assumptions of an envisioned operating environment (application) explicit. A desired security-convenience trade-off is specified through detection costs assigned to erroneous system decisions and prior probabilities assigned to the commonality of targets, non-targets, and spoofing attacks. The ASV threshold is set to the EER point on bona fide samples. Following the notations and parameter constraints in [34], we assign costs $C_{\text{misc}} = 1$, $C_{t} = C_{\text{t,asvspoof}} = 10$, and priors $\pi_{\text{t,spoof}} = 0.50$, $\pi_{\text{t,non}} = (1 - \pi_{\text{t,spoof}}) \times 0.99 \approx 0.495$, $\pi_{\text{t,asvspoof}} = (1 - \pi_{\text{t,spoof}}) \times 0.01 \approx 0.005$. This is representative of a high-security user authentication ap-

Table 6: Coefficients (p-values) of multiple linear regression models that use ASR WER, MOSNet predictions, ASV EER, and countermeasure EER as inputs. Top: subjective scores from English listeners. Bottom: subjective scores from Japanese listeners.

| Subjective score (ENG) | Intercept | MOSNet (asvspoof19) | ASR WER (%) | ASV EER (%) | Countermeasure EER (%) | Multiple R-Squared | Adjusted R-squared | Significance F |
|------------------------|-----------|---------------------|-------------|-------------|------------------------|-------------------|-------------------|----------------|
| Task 1 MOS             | 1.713     | 0.258               | -0.021      | 0.024       | -0.002                 | 0.92              | 0.81              | <0.001         |
| Task 1 SIM             | 1.696     | 0.062               | -0.003      | 0.026       | 1.006                  | 0.92              | 0.83              | <0.001         |
| Task 2 MOS             | 0.619     | 0.724               | -0.021      | 0.014       | 0.003                  | 0.88              | 0.74              | <0.001         |
| Task 2 SIM             | 1.782     | 0.038               | -0.002      | 0.026       | 0.002                  | 0.91              | 0.80              | <0.001         |

| Subjective score (JPN) | Intercept | MOSNet (asvspoof19) | ASR WER (%) | ASV EER (%) | Countermeasure EER (%) | Multiple R-Squared | Adjusted R-squared | Significance F |
|------------------------|-----------|---------------------|-------------|-------------|------------------------|-------------------|-------------------|----------------|
| Task 1 MOS             | 1.024     | 0.364               | -0.011      | 0.024       | -0.002                 | 0.89              | 0.76              | <0.001         |
| Task 1 SIM             | 1.439     | 0.132               | -0.001      | 0.024       | 0.006                  | 0.91              | 0.81              | <0.001         |
| Task 2 MOS             | -0.040    | 0.833               | -0.014      | 0.014       | 0.006                  | 0.85              | 0.68              | <0.001         |
| Task 2 SIM             | 1.776     | 0.059               | -0.003      | 0.023       | 0.004                  | 0.88              | 0.74              | <0.001         |
Table 7: Minimum t-DCF for each system of VCC 2020. Red cells indicate top-5 systems for each task.

| System | Task 1 | Task 2 | System | Task 1 | Task 2 |
|--------|--------|--------|--------|--------|--------|
| T01    | 0.73542 | –      | T18    | 0.70372 | 0.81145 |
| T02    | 0.85274 | 0.70888 | T19    | 0.8743  | 0.90471 |
| T03    | 0.01467 | 0.01467 | T20    | 0.85051 | 0.77249 |
| T04    | 0.88342 | 0.8617   | T21    | 0.86755 |        |
| T05    | –       | 0.69094 | T22    | 0.86204 | 0.93512 |
| T06    | 1.00000 | 0.72722 | T23    | 0.8297  | 0.9037  |
| T07    | 0.87227 | 0.9033  | T24    | 0.76482 | 0.79092 |
| T08    | 1.00000 | 1.00000 | T25    | 0.85402 | 0.85048 |
| T09    | 0.25987 | 0.29213 | T26    | 0.71041 | 0.53263 |
| T10    | 0.87126 | 0.91252 | T27    | 0.80151 | 0.84287 |
| T11    | 0.87531 | 0.88646 | T28    | 0.91214 | 0.82598 |
| T12    | 1.00000 | 0.84693 | T29    | 0.83375 | 0.87311 |
| T13    | 0.88646 | 0.79685 | T30    | 0.04508 | 0.09695 |
| T14    | 0.91708 | –       | T31    | 0.84069 | 0.70379 |
| T15    | –       | 0.8805  | T32    | 0.80942 | 0.76208 |
| T16    | 0.87633 | 0.88818 | T33    | 0.78905 | 0.83375 |
| T17    | 0.87734 | –       | –      | –      | –      |

Table 8: Details of top-performing VC systems in terms of minimum t-DCF as a spoofing threat.

| Task 1 | System | Task 2 |
|--------|--------|--------|
| Team ID | VC model | Vocoder |
| T06    | StarGAN | WORLD  |
| T08    | VTLN + Spectral differential | WORLD  |
| T12    | ADAGAN  | AHRcoder |
| T14    | One-shot VC | NSF    |
| T28    | Tacotron | WaveRNN |

Note: The mathematical properties of t-DCF [34] imply that if the spoof false acceptance rate (SAR) of the ASV system is 0, the optimal countermeasure is no countermeasure. Interested readers are directed to [34, Eq. (10)], where $C_2 = 0$ whenever $P_{fa,spoof} = 0$. A CM that minimizes Eq. (10) in this case must use a detection threshold $\gamma^{th} = -\infty$, aka ‘accept everything’ or ‘no countermeasure’.

3The mathematical properties of t-DCF [34] imply that if the spoof false acceptance rate (SAR) of the ASV system is 0, the optimal countermeasure is no countermeasure. Interested readers are directed to [34, Eq. (10)], where $C_2 = 0$ whenever $P_{fa,spoof} = 0$. A CM that minimizes Eq. (10) in this case must use a detection threshold $\gamma^{th} = -\infty$, aka ‘accept everything’ or ‘no countermeasure’.

6.2. Discussion

Do the VC methods featured in the VCC 2020 impose a spoofing risk? Yes, they certainly do. One useful reference is the ASV performance on natural human samples—more specifically, EER on target-nontarget discrimination. These EERs were 0.50% and 0.80% for Tasks 1 and 2, respectively. Most VC systems increased these numbers, most of them substantially. This is not news to the ASV community as such. Our spoofing performance assessment through t-DCF clearly highlights the importance of improving both ASV and CM technology. The VCC 2020 not only featured a battery of new VC techniques but also facilitated an initial CM performance benchmarking on a new type of multilingual data. Further research is needed to analyze the combined effect of VC methods and language. Substantial future work remains in improving the generalizability of CM techniques across diverse data conditions.

7. Conclusion

This work summarizes the predictions of subjective ratings and spoofing assessments with objective assessments performed at the latest VCC 2020. We considered five different objective assessments based on ASV, neural speaker embedding, spoofing countermeasure, predicted MOS, and ASR. The correlations of objective assessments computed with subjective ratings indicate that the ASV, neural speaker embedding, and ASR had high correlations, which suggests the possibility of predicting the subjective ratings. Further, we found that the ASV and ASR results were more effective than the predicted MOS and spoofing countermeasure for predicting the subjective test results using multiple linear regression models. This indicates the potential shift toward relying on the objective assessments over tedious listening tests for large-scale evaluations in the future. We performed further spoofing assessment on the submissions and identified
the VC methods with a potentially high threat. However, this topic deserves future exploration as the performance highly depends on the coverage of the various VC methods included in the training data.

8. Acknowledgements

The authors thank Ms. Jennifer Williams of the University of Edinburgh for kindly providing a MOSNet model fine-tuned using the ASVspoof 2019 data. This work was partially supported by JST CREST Grants (JPMJCR18A6, VoicePersonae project, and JPMJCR19A3, CoAugmentation project), Japan, MEXT KAKENHI Grants (16H06302, 17H04687, 17H06101, 18H04120, 18H04112, 18KT0051, 19K24373), Japan, the National Natural Science Foundation of China (Grant No. 61871358), Programmatic Grant No. A1687b0033 from the Singapore Governments Research, Innovation and Enterprise 2020 plan (Advanced Manufacturing and Engineering domain), and the Academy of Finland (project no. 309629).

9. References

[1] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Toda, T. Kinnunen, and Z. Ling, “Voice conversion challenge 2020 intra-lingual semiparallel and cross-lingual voice conversion,” in ISCA Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020, 2020, pp. XXX–XXX.

[2] T. Yoshimura, G. Eje Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, “A hierarchical predictor of synthetic speech naturalness using neural networks,” in Interspeech 2016, 2016, pp. 342–346.

[3] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. W. Wilson, R. A. Saurous, and D. Sculley, “AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech,” in NIPS End-to-end Learning for Speech and Audio Processing Workshop, 2016.

[4] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, “Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM,” in Interspeech 2018, 2018, pp. 1873–1877.

[5] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, “Non-intrusive speech quality assessment using neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019, 2019, pp. 631–635.

[6] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “MOSNet: Deep learning-based objective assessment for voice conversion,” in Interspeech 2019, 2019, pp. 1541–1545.

[7] J. Williams, J. Rownicka, P. Oplustil, and S. King, “Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis,” in Odyssey 2020 The Speaker and Language Recognition Workshop, 2020, pp. 222–229.

[8] Y. Choi, Y. Jung, and H. Kim, “Deep MOS predictor for synthetic speech using cluster-based modeling,” arXiv preprint arXiv:2008.03710, 2020.

[9] ——, “Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification,” arXiv preprint arXiv:2007.08267, 2020.

[10] B. T. Meyer, B. Kollmeier, and J. Ooster, “Autonomous measurement of speech intelligibility utilizing automatic speech recognition,” in Interspeech 2015, 2015, pp. 2982–2986.

[11] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” Speech Communication, vol. 66, pp. 130 – 153, 2015.

[12] R. K. Das, X. Tian, T. Kinnunen, and H. Li, “The attacker’s perspective on automatic speaker verification: An overview,” in Interspeech 2020, 2020.

[13] https://kaldi-asr.org/models/m7/.

[14] G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlos, “STC antispoofing systems for the ASVspoof2019 challenge,” in Interspeech 2019, 2019, pp. 1033–1037.

[15] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-end attention-based large vocabulary speech recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, 2016, pp. 4945–4949.

[16] A. W. Rix, J. G. Beersend, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2001, vol. 2, 2001, pp. 749–752.

[17] J. G. Beersend, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, “Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part Itemporal alignment,” Journal of the Audio Engineering Society, vol. 61, no. 6, pp. 366–384, 2013.

[18] L. Malfait, J. Berger, and M. Kastner, “P. 5638 ITU-T standard for single-ended speech quality assessment,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 1924–1934, 2006.

[19] S. King, “Measuring a decade of progress in text-to-speech,” Lorrains, vol. 1, no. 1, p. 106, 2014.

[20] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016,” in Interspeech 2016, 2016, pp. 1632–1636.

[21] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” in Odyssey 2018 The Speaker and Language Recognition Workshop, 2018, pp. 195–202.

[22] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Siahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvola, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Stein, D. Matrouf, J.-F. Bonastre, A. Govenier, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, “Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Computer Speech and Language, vol. 64, p. 101114, 2020.

[23] Huadi Zheng, W. Cai, Tianyan Zhou, Shilei Zhang, and M. Li, “Text-independent voice conversion using deep neural network based phonetic level features,” in International Conference on Pattern Recognition (ICPR), Dec 2016, pp. 2872–2877.

[24] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, “Phonetic posteriorgrams for many-to-one voice conversion without parallel data training,” in IEEE International Conference on Multimedia and Expo (ICME) 2016, 2016, pp. 1–6.

[25] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, “Voice conversion using sequence-to-sequence learning of context posterior probabilities,” in Interspeech 2017, 2017.

[26] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, “WaveNet vocoder with limited training data for voice conversion,” in Interspeech 2018, 2018, pp. 1983–1987.

[27] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, “Sequence-to-sequence acoustic modeling for voice conversion,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 3, pp. 631–644, 2019.

[28] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, “Improving sequence-to-sequence acoustic modeling by adding text-supervision,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019, 2019, pp. 6785–6789.
[29] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in *Interspeech 2017*, F. Lacerda, Ed., 2017, pp. 999–1003.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *IEEE Workshop on Automatic Speech Recognition and Understanding 2011*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[31] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” *Computer Speech and Language*, vol. 60, p. 101027, 2020.

[32] M. Sahidullah, T. Kinnunen, and C. Hanilçi, “A comparison of features for synthetic speech detection,” in *Interspeech 2015*, 2015, pp. 2087–2091.

[33] Z. Wu, R. K. Das, J. Yang, and H. Li, “Light convolutional neural network with feature genuinization for detection of synthetic speech attacks,” in *Interspeech 2020*, 2020.

[34] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2195–2210, 2020.
A. Results of Individual Metrics

Here, we show the rankings and comparison of every team based on the various objective measures for both tasks discussed in Section 4.

Figure 1: Summary of ASV-based speaker similarity assessment for Task 1.

Figure 2: Summary of ASV-based speaker similarity assessment for Task 2.

Figure 3: Summary of cosine distance-based neural speaker embedding for Task 1.
**Task 2 Cosine Distance Performance**

Figure 4: *Summary of cosine distance-based neural speaker embedding for Task 2.*

**Task 1 Countermeasure Performance**

Figure 5: *Summary of spoofing countermeasure EER (%) for Task 1.*

**Task 2 Countermeasure Performance**

Figure 6: *Summary of spoofing countermeasure EER (%) for Task 2.*
Figure 7: *MOSNet predictions of all systems in Task 1.*

Figure 8: *MOSNet predictions of all systems in Task 2.*

Figure 9: *Summary of ASR WER (%) for Task 1.*

Figure 10: *Summary of ASR WER (%) for Task 2.*
B. Objective Evaluation Results and Scatter Plots

The analysis presented here shows the correlation scatter plots of various objective measures against the subjective MOS and speaker similarity for both English and Japanese listeners. The Pearson correlation coefficients along with the $p$-values corresponding to Figures 11-24 have been presented in Table 4.

Figure 11: Scatter plots for ASV EER (%) with subjective MOS.

Figure 12: Scatter plots for ASV EER (%) with subjective speaker similarity.
Figure 13: Scatter plots for ASV Pfa (%) with subjective MOS.

Figure 14: Scatter plots for ASV Pfa (%) with subjective speaker similarity.
Figure 15: Scatter plots for speaker embedding cosine distance with subjective MOS.

Figure 16: Scatter plots for speaker embedding cosine distance with subjective speaker similarity.
Figure 17: Scatter plots for spoofing countermeasure EER (%) with subjective MOS.

Figure 18: Scatter plots for spoofing countermeasure EER (%) with subjective speaker similarity.
Figure 19: Scatter plots for MOSNet (vcc18) predictions with subjective MOS.

Figure 20: Scatter plots for MOSNet (vcc18) predictions with subjective speaker similarity.
Figure 21: Scatter plots for MOSNet (asvspoof19) predictions with subjective MOS.

Figure 22: Scatter plots for MOSNet (asvspoof19) predictions with subjective speaker similarity.
Figure 23: Scatter plots for ASR WER (%) with subjective MOS.

Figure 24: Scatter plots for ASR WER (%) with subjective speaker similarity.
C. Objective Evaluation Results by Target Speaker Language

Here, we perform an analysis on the effect of target speaker language on Task 2 of VCC 2020 for different objective measures on all the systems. The correlation of each objective measure with the subjective test for the language pairs is presented in Table 5. In the figures, “Fin”, “Ger”, and “Man” stand for Finnish, German, and Mandarin, respectively.

Figure 25: ASV performance in EER (%) of various teams with different language pair analysis in Task 2 of VCC 2020.

Figure 26: ASV performance in Pfa (%) of various teams with different language pair analysis in Task 2 of VCC 2020.

Figure 27: Neural speaker embedding cosine similarity of various teams with different language pair analysis in Task 2 of VCC 2020.
Figure 28: Spoofing countermeasure performance in EER (%) of various teams with different language pair analysis in Task 2 of VCC 2020.

Figure 29: MOSNet (vcc18) predictions of all systems for different target speaker languages on Task 2 of VCC 2020.

Figure 30: MOSNet (asvspoof19) predictions of all systems for different target speaker languages on Task 2 of VCC 2020.

Figure 31: ASR performance in WER (%) of various teams with different language pair analysis in Task 2 of VCC 2020.