THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Predicting pairwise preferences between TTS audio stimuli using parallel ratings data and anti-symmetric twin neural networks

Predicting pairwise preferences between TTS audio stimuli using parallel ratings data and anti-symmetric twin neural networks

Cassia Valentini-Botinhao[1,2], Manuel Sam Ribeiro[1], Oliver Watts[2],*
Korin Richmond[1], Gustav Eje Henter[3]

[1]The Centre for Speech Technology Research, University of Edinburgh, UK
[2]SpeakUnique Ltd., UK
[3]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

cvbotinh@inf.ed.ac.uk

Abstract

Automatically predicting the outcome of subjective listening tests is a challenging task. Ratings may vary from person to person even if preferences are consistent across listeners. While previous work has focused on predicting listeners' ratings (mean opinion scores) of individual stimuli, we focus on the simpler task of predicting subjective preference given two speech stimuli for the same text. We propose a model based on anti-symmetric twin neural networks, trained on pairs of waveforms and their corresponding preference scores. We explore both attention and recurrent neural nets to account for the fact that stimuli in a pair are not time aligned. To obtain a large training set we convert listeners' ratings from MUSHRA tests to values that reflect how often one stimulus in the pair was rated higher than the other. Specifically, we evaluate performance on data obtained from twelve MUSHRA evaluations conducted over five years, containing different TTS systems, built from data of different speakers. Our results compare favourably to a state-of-the-art model trained to predict MOS scores.

Index Terms: Preference prediction, text-to-speech, twin neural networks, MUSHRA

1. Introduction

Human ratings are the gold standard in evaluating speech generation technologies. Participants are asked to listen to speech stimuli and rate them in isolation or in context. Ratings are subjective, varying from person to person, even more so in cases where the listener is given less context and training. The most common type of evaluation is the MOS test [1]. MOS scores are collected from listeners hearing each utterance in isolation, relying on the fact that every person has an internal reference for what is highly natural (score 5) and unnatural (score 1). Even when relative preferences are consistent across listeners, MOS values may vary from person to person. Paired comparison tests (so-called 'AB tests') exhibit a much lower variance in their responses, which in turn makes statistical significance easier to detect [2, 3]. Although it is possible to retrieve ordinal rankings from AB tests [4], MOS (and to some extent MUSHRA tests [5]) are presumably more widely used as their results can directly be used to rank systems in a single scale. In a MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) test, participants are asked to score systems on a scale from 1–100 by listening to stimuli for the same text side-by-side alongside a high quality reference. Unlike MOS tests, where ratings are given in isolation, MUSHRA test participants are performing a multiple comparison test. This design makes it more sensitive to small differences between stimuli than MOS tests.

Obtaining reliable results from subjective tests requires careful design and a considerable number of listeners [6]. An easier alternative to human evaluation is promised by so-called objective measures. They can be computed for a given waveform stimulus automatically and are deemed objective as they rely on the same computation no matter who performs it. Although each measure is deterministic, different measures might give drastically different results. The accuracy of a metric (how closely it correlates with human ratings) will most likely depend on the particular characteristics of the stimuli (the speaker, the listener, the linguistic content, the quality of the recording, etc). Designing a measure that works well in every condition is a very challenging task. While measures that are based on auditory models fail to generalise to conditions where the model alone cannot explain human scores, data-driven measures struggle on conditions that were unseen during training.

Most objective measures were in fact designed to assess natural speech that has been corrupted by a telecommunication channel and/or environmental noise (notably PESQ [7], POLQA [8] and STOI [9]) rather than artificially generated speech. There has been a growing interest however in metrics for synthesised speech (speech generated by text-to-speech (TTS) systems or voice conversion). Most prominent among these are machine learning models trained on paired audio rating data. These include the models in [10] which uses the data from Blizzard Challenge 2012 (220 datapoints); the model in [11], trained on data from 6 years of the Blizzard Challenge [12] (3,324 datapoints); AutoMOS, trained on an extensive proprietary dataset of human scores (47,320 datapoints) [13]; MOS-Net [14], first trained on voice conversion data from the Voice Conversion Challenge (13,580 datapoints) and later on both TTS and voice conversion data from one of the tracks of the ASVspoof Challenge 2019 (number of datapoints not reported) [15]; and more recently the model proposed in [16] trained on 8 years of Blizzard Challenge data, 2 years of Voice Conversion Challenge plus augmented data generated using POLQA (number of datapoints not reported). All these models are trained to predict MOS ratings for a single waveform. Experience shows that predicting MOS is a very challenging task (esp. for unseen speakers, listeners and systems in the out-of-domain setting [17]), requiring large amounts of training data.

Although AB scores are more consistent and potentially easier to predict, we know of only one study that has looked into predicting AB scores derived from listening tests [18]. They proposed a neural network that is trained to predict whether or not two stimuli are perceived as different in listening test re-

---

sponses. To obtain a balanced training set, they designed an adaptive listening test that gradually distorts speech using, e.g., additive noise, reverberation and compression. It is not clear how one would gradually distort audio to obtain similar data for text-to-speech. In their more recent work [19] the authors incorporate contrastive learning and multi-dimensional learning to encourage the model to learn 'content' invariant representations that better generalise to different speakers and sentences.

In this paper we propose *PrefNet*, a neural network trained to predict listener preference given two speech stimuli. In order to create a sizeable set of preference data to train PrefNet we propose converting MUSHRA scores derived from several listening evaluations to pairwise preference scores. We evaluate our system using unseen data (different voices and synthesisers) and compare it to MOSNet [14]. Our code, pre-trained models and data are publicly available.[1]

2. Method

2.1. Preference scores

To convert MUSHRA scores into pairwise preference scores we compared the scores of every pair of stimuli belonging to the same MUSHRA screen as shown to the left in Fig. 1.

This shows which out of two stimuli (different systems speaking the same utterance) a listener preferred. The numerical difference between the scores is not taken into account as we assume every participant might use the 0–100 scale differently. The final preference scores were calculated on a per screen basis (i.e., per waveform) as the average of preference scores across participants (right side of Fig. 1). The strength of the preference is therefore measured in terms of how many participants rated one stimulus above the other, rather than the difference in average score between the two stimuli.

2.2. Preference prediction

Fig. 2 shows the general architecture of PrefNet. Each input waveform is processed by an encoder network made of a mel spectogram extraction layer (with fixed coefficients) followed by several convolutional layers (which are shared across the different inputs). The variable length encoded representations are then processed by an alignment/scoring mechanism which returns a vector $d(\cdot)$ that essentially subtracts the encoded version of X_B from the encoded version of X_A. That vector is then processed by a fully connected network $f(\cdot)$ and a sigmoid activation $\sigma(\cdot)$ as follows:

$$\Phi(X_A, X_B) = \sigma(f(d(X_A, X_B)) - f(-d(X_A, X_B))) \quad (1)$$

where $X_A \in \mathbb{R}^{N_A \times 1}$ and $X_B \in \mathbb{R}^{N_B \times 1}$, N_A and N_B are the number of mel spectogram frames for each input.
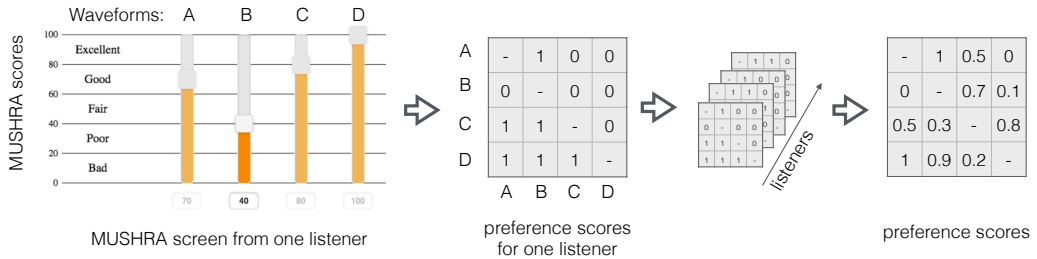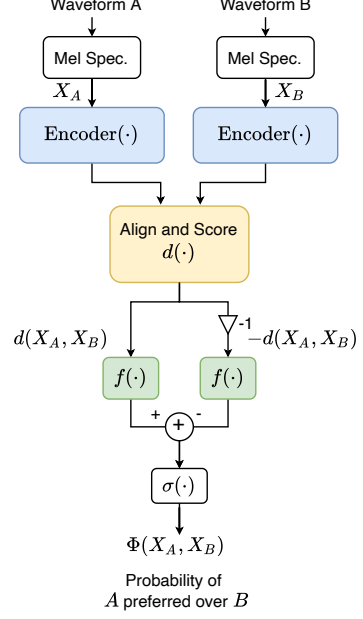
---
[1] https://github.com/cassiavb/PrefNet

Figure 2: *The general architecture of PrefNet. Same-colour boxes denote identical networks with the same weights.*

PrefNet is trained to predict the (empirical) probability of how often waveform A was preferred over waveform B. The purpose of Eq. (1) is to ensure that the predicted probabilities are consistent when the input waveforms are swapped: $\Phi(X_B, X_A) = 1 - \Phi(X_A, X_B)$. This should lead to more data-efficient modelling, since this property is built in and does not have to be learned.

Architectures where the exact same neural network is applied to multiple inputs are known as *twin neural networks* (or Siamese neural networks). These are often employed in order to learn functions that are symmetric in the input arguments, e.g., similarity scores. The design of PrefNet instead leverages a twin setup to learn a function that is anti-symmetric in the two input stimuli. (Specifically, the output, prior to the final sigmoid, is multiplied by -1 if the two inputs change places.) We therefore call our setup an *anti-symmetric twin neural network*.

We propose two versions of PrefNet that vary in terms of how alignment and scoring is performed.

2.2.1. Attention-based scoring

We investigate an attention-like mechanism [29] to align and score the distance between input stimuli. The encoder network of the attention-based model predicts two sets of representations

Figure 1: *Converting same-screen MUSHRA scores (left) to preference scores (right).*

	Large set	Eval$_1$	Eval$_2$	Eval$_3$	Eval$_4$	Eval$_5$	Eval$_6$	Eval$_7$
Stimuli pairs	17,047	2,160	1,440	800	800	800	3,300	720
System pairs	282	36	36	10	10	10	55	10
Systems	76	9	9	5	5	5	11	5
Speech corpus	Hurricane [20]	Blizzard 2013 [21]	Blizzard 2013 [21]	Blizzard 2013 [21]	Blizzard 2013 [21]	Blizzard 2013 [21]	LJ Speech [22]	Blizzard 2013 [21]
TTS systems	HTS [23], Multisyn [24], vocoded [25], Merlin [26]	DCTTS [27]	DCTTS [27]	DCTTS [27]	DCTTS [27]	DCTTS [27]	Merlin [26], DCTTS [27]	DCTTS [27], Tacotron [28]

Table 1: *Overview of the data used in the evaluation. The 'Large set' contains data from 12 different evaluations.*

K and V for each input:

$$K_A, V_A = H(\text{Encoder}(X_A)) \qquad (2)$$
$$K_B, V_B = H(\text{Encoder}(X_B)) \qquad (3)$$

where $H(\cdot)$ is an affine layer, $K_A \in \mathbb{R}^{N_A \times D}$, $K_B \in \mathbb{R}^{N_B \times D}$, $V_A \in \mathbb{R}^{N_A \times E}$ and $V_B \in \mathbb{R}^{N_B \times E}$. The attention matrix is then computed as follows:

$$W = \operatorname*{softmax}_{N_A, N_B}(K_A K_B^T) \qquad (4)$$

where $W \in \mathbb{R}^{N_A \times N_B}$. Note that the softmax operation is performed across both rows and columns, i.e., all elements of W collectively sum to one. The distance $d(\cdot) \in \mathbb{R}^{E \times 1}$ between the two inputs is derived as:

$$d(X_A, X_B) = \sum_{t_A, t_B=1}^{t_A=N_A, t_B=N_B} (V_A[t_A, :] - V_B[t_B, :]) W[t_A, t_B]. \qquad (5)$$

2.2.2. GRU-based scoring

Rather than comparing the input signals frame by frame, as the attention-based system does, our other system variants process each variable length representation with a shared GRU layer and convert the processed sequence into a fixed length representation. The scoring vector $d(\cdot)$ is then computed as follows:

$$d(X_A, X_B) = g(\text{GRU}(\text{Encoder}(X_A))) \\ - g(\text{GRU}(\text{Encoder}(X_B))) \qquad (6)$$

where $d \in \mathbb{R}^{E \times 1}$ and $g(\cdot)$ represents the operation that extracts a fixed length representation from the GRU sequence, like from instance its last output.

3. Evaluation

3.1. Data

Table 1 presents details on the data we used for training and testing. The first column ('Large set') describes data collected from 12 different lab-based MUSHRA tests over a period of 5 years (2014–2019). This includes the evaluations described in [30, 31, 32, 33, 25, 34, 35], among others. The tests involve stimuli of natural and vocoded speech, but with a focus on a variety of TTS systems, including HMM-based systems build with HTS [23], unit selection built with Multisyn [24] and early DNN systems built with Merlin [26]. All systems in these evaluations were created using the same speech dataset, the Hurricane Challenge corpus (male British speaker) [20].

Test fold:	Fold 1	Fold 2	Fold 3	Fold 4	Average
Training folds					
# pairs	12,281	10,237	15,647	12,976	-
Attention	70.0	81.3	70.8	67.5	72.4
GRU$_1$	91.2	90.4	90.6	90.0	90.5
GRU$_2$	91.9	92.3	91.8	90.4	91.6
GRU$_3$	**92.7**	**92.4**	92.2	90.5	**91.9**
GRU$_4$	92.5	91.8	**92.6**	90.7	**91.9**
Test fold					
# pairs	4,766	6,810	1,400	4,071	-
Attention	60.0	66.6	64.8	63.6	63.8
GRU$_1$	68.7	70.8	73.6	62.2	68.8
GRU$_2$	72.9	73.4	72.4	65.2	71.0
GRU$_3$	75.3	**75.2**	75.1	70.1	73.9
GRU$_4$	**76.9**	74.2	**77.1**	**71.5**	**74.9**

Table 2: *Prediction accuracy (%) of different folds (columns) on the training data (top rows) and on the test data (bottom rows).*

The other columns in Table 1 describe data from more recent lab-based MUSHRA evaluations, containing sequence-to-sequence TTS models like DCTTS [27] and Tacotron [28]. Eval$_6$ described in [36] involved voices built with Merlin and DCTTS using the LJ Speech dataset (US female speaker) [22]. All other evaluations contained voices based on the Blizzard Challenge 2013 corpus (US female speaker) [21].

All audio stimuli were resampled to 16 kHz. We extracted a 64-dimensional log-magnitude mel spectrogram representation using a window of width 512 and 12.5 ms frame hop.

3.2. Models

We performed a hyperparameter sweep on the 'Large set' in Table 1 using multifold cross-validation. The best attention-based model we found was made of four convolutional layers of width three (the first layer of dilation three, the others one) over 64 channels (D=32 and E=32). The encoder of the best GRU-based model was made of two convolutional layers of kernel width 9 over 64 channels. The GRU contained 64 units (E=64). For all models the fully connected network $f(\cdot)$ had only one layer. All models were trained to predict relative preferences (probabilities) using the mean squared error loss (also known as the Brier score). Training ran for 50 epochs with early stopping based on a random selection of 10% of the training data for validation, using the Adam optimiser with a learning rate of 0.001. The pairs of input sequences were batched according to the size of the longer waveform in each pair.

We present results of four GRU-based models. GRU$_1$ is a variant without the anti-symmetric design whose output is instead given by $\sigma(f(g(\text{Encoder}(X_A)) - g(\text{Encoder}(X_B))))$. GRU$_2$ is the anti-symmetric version proposed in Eqs. (1) and (6) where $g(\cdot)$ extracts the final output of the GRU. GRU$_3$ obeys the same anti-symmetric architecture but rather than retrieving

Train \ Test	Eval$_1$	Eval$_2$	Eval$_3$	Eval$_4$	Eval$_5$	Eval$_6$	Eval$_7$
PrefNet							
Large set	48.8 / 55.6	46.6 / 44.4	55.4 / 60.0	52.8 / 60.0	57.5 / 80.0	56.8 / 61.8	**68.2 / 100**
Eval$_1$	-	**72.9 / 83.3**	**75.5 / 80.0**	67.8 / 70.0	**84.8 / 90.0**	**66.3 / 65.5**	61.4 / 60.0
Large set and Eval$_1$ (FT)	-	68.8/ **88.9**	69.9 / **80.0**	63.2 / **80.0**	83.6 / **100**	65.5 / 70.9	63.5 / 70.0
Eval$_{1-5}$	-	-	-	-	-	62.4 / 61.8	65.6 / 50.0
Large set and Eval$_{1-5}$ (FT)	-	-	-	-	-	66.1 / **72.7**	63.2 / 50.0
MOSNet							
Voice conversion data (VC) [14]	51.2 / 50.0	49.2 / 33.3	68.6 / **80.0**	45.9 / 60.0	51.0 / 70.0	65.1 / 67.3	51.5 / 30.0
Anti-spoofing data [15] (TTS, TTS+VC, VC)	**68.8 / 86.1**	61.5 / 69.4	54.1 / 60.0	44.9 / 40.0	65.8 / 80.0	43.4 / 40.0	31.4 / 20.0

Table 3: *Prediction accuracy (%) at stimuli / system level. FT stands for fine-tuned.*

the final output, the average representation is obtained. GRU$_4$ is a bi-directional version of GRU$_3$.

3.3. Similar conditions: same speaker

To compare variants of the PrefNet architecture, we performed a multifold cross-validation evaluation using the 'Large set' data. We divided the 12 evaluations into four folds ordering the evaluations chronologically, so that different folds are less likely to contain the same systems. (For instance, the first and second fold contained only HMM-based TTS, while only the third and fourth folds had DNN-based voices.) Each architecture was trained four times, varying which fold was held out from training. The prediction accuracy (calculated at a sentence level across all sentences) is reported in Table 2.

We see that accuracy on the training data (upper rows) is relatively high, particularly for the GRU-based models. Performance drops considerably when evaluating on unseen material, i.e., the fold that has been held out (lower rows). Overall we can see that the GRU-based models performs better than the attention-based one. GRU$_2$ scores higher than GRU$_1$ indicating that the anti-symmetric architecture is helpful. We observed that introducing it improved not only accuracy, but also model convergence. Taking the average over the GRU's output sequence (rather than using its last element) seems to further improve results, as well as using bidirectional GRUs. The remaining experiments focus on the GRU$_4$ variant of PrefNet.

3.4. Unseen conditions: different speakers and systems

To evaluate how well PrefNet predicts the preference between more recent state-of-the-art TTS models we extended the evaluation to the data from evaluations Eval$_1$ to Eval$_7$, with several different training and test data configurations. The results are presented in Table 3. Rows refer to training material and columns to the test set. We present scores at both stimulus level and system level. For system-level results we calculated the accuracy across all sentences of each system pair.

As a baseline, we compare the various PrefNet results to results obtained from pre-trained MOSNet models, one trained on voice conversion (VC) data [14] and one trained on both VC and TTS data [15] including HMM-based systems, Merlin and Tacotron trained on the VCTK corpus (multi-speaker English language dataset) [37]. These represent the state of the art in publicly available models for MOS score prediction on synthetic speech. The MOS scores predicted by MOSNet were converted to pairwise preferences by checking which of two paired stimuli had the higher predicted MOS score.

The results in Table 3 show that the model trained on the 'Large set' (top row) performs relatively well on Eval$_7$ but not on the other evaluations where the stimulus-level perform-

ance is close to chance rate. The model trained on Eval$_1$ alone (second row) highlights the importance of representative training data, since it performs well on similar evaluations (i.e., Eval$_{2-5}$) but less so on evaluations with different speakers and systems (Eval$_6$ and Eval$_7$). Fine tuning the model trained on the 'Large set' with data from Eval$_1$ (third row) seems to improve results on Eval$_{2-5}$ but to the detriment of performance on Eval$_7$. Training with more quantities of similar data (fourth and fifth row) does not seem to improve performance on Eval$_6$. System-level performance is generally better than stimulus-level performance for all systems. System-level accuracy seems to be higher for the fine-tuned models.

The MOSNet models generally perform worse than most PrefNet models, with performance varying considerably depending on the specific test set. Poor performance could to some extent be explained by the fact that these models were trained on data that is potentially more dissimilar to the test set than PrefNet training data. In several cases, MOSNet performance is below chance rate, with Eval$_7$ being the most challenging to predict. Interestingly, the MOSNet model trained on TTS data was not always the best. A possible explanation is that it used less training data than the VC-only model.

4. Conclusions

This paper has introduced PrefNet, an approach to predicting pairwise preferences between synthetic speech stimuli. We demonstrated how data from side-by-side evaluations using numerical ratings can be leveraged to create training data for pairwise preference prediction. We empirically investigated several architectures and described a design rooted in twin neural nets that ensures consistent pairwise preferences if the order of the inputs is reversed. Results showed that GRU-based architectures outperformed those using attention to align and score stimulus pairs, and that our anti-symmetric network design also improved accuracy. PrefNet outperforms MOSNet (a state-of-the-art model trained to predict MOS scores, from which implicit pairwise preferences can be derived) in most conditions but the performance of both models varies considerably depending on training data and test data. The performance of the model trained on the larger set of listening tests (including HMM and DNN systems) improves when data from more recent (and thus more similar) systems are included in training. Future work includes training with both stimulus-level and system-level preferences and increasing the amount of training data by converting MOS (as well as MUSHRA) scores to preferences.

5. References

[1] International Telecommunication Union, Telecommunication Standardization Sector, "Methods for subjective determination of transmission quality," ITU Recommendation ITU-T P.800, 1996.

[2] Y. Vazquez-Alvarez and M. Huckvale, "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems." in *Proc. ICSLP*, 2002, pp. 329–332.

[3] S. Kiritchenko and S. Mohammad, "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," in *Proc. ACL*, 2017, pp. 465–470.

[4] S. Shirali-Shahreza and G. Penn, "MOS naturalness and the quest for human-like speech," in *Proc. SLT Workshop*, 2018, pp. 346–352.

[5] International Telecommunication Union, Radiocommunication Sector, "Method for the subjective assessment of intermediate quality levels of coding systems," ITU Recommendation ITU-R BS.1534-3, 2015.

[6] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proc. Interspeech*, 2015, pp. 3476–3480.

[7] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, Salt Lake City, USA, May 2001, pp. 749–752.

[8] J. Beerends, 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 I-temporal alignment," *J. Aud. Eng. Soc.*, vol. 61, pp. 366–384, 06 2013.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, USA, March 2010, pp. 4214–4217.

[10] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks – Blizzard Challenge 2012," in *Proc. Blizzard Challenge Workshop*, 2012.

[11] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks," in *Proc. Interspeech*, 2016, pp. 342–346.

[12] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, p. e006, 2014.

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

[14] 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 *Proc. Interspeech*, 2019, pp. 1541–1545.

[15] 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 *Proc. Speaker Odyssey*, 2020.

[16] G. Mittag and S. Möller, "Deep learning based assessment of synthetic speech naturalness," in *Proc. Interspeech*, 2020.

[17] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," 2022. [Online]. Available: https://arxiv.org/abs/2203.11389

[18] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual audio metric learned from just noticeable differences," in *Proc. Interspeech*, 2020, pp. 2852–2856.

[19] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "CDPAM: Contrastive learning for perceptual audio similarity," in *Proc. ICASSP*, Jun. 2021.

[20] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus," https://doi.org/10.7488/ds/140, 2013.

[21] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Proc. Blizzard Challenge Workshop*, 2013.

[22] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[23] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW*, 2007, pp. 294–299.

[24] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Comm.*, vol. 49, no. 4, pp. 317–330, 2007.

[25] F. Espic Calderón, C. Valentini Botinhao, and S. King, "Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis," in *Proc. Interspeech*, 2017.

[26] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. SSW*, Sept. 2016, pp. 218–223.

[27] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. ICASSP*, 2018, pp. 4784–4788.

[28] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. ICML*, 2018, pp. 4693–4702.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[30] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis." in *Proc. ICASSP*, 2015.

[31] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context HMM synthesis," in *Proc. Interspeech*, 2015.

[32] T. Merritt, R. A. J. Clark, Z. Wu, J. Yamagishi, and S. King, "Deep neural network-guided unit selection synthesis," in *Proc. ICASSP*, 2016.

[33] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: Where do the improvements come from?" in *Proc. ICASSP*, 2016.

[34] O. Watts, C. Valentini Botinhao, F. Espic Calderón, and S. King, "Exemplar-based speech waveform generation," in *Proc. Interspeech*, 2018.

[35] Z. Malisz, G. E. Henter, C. Valentini-Botinhao, O. Watts, J. Beskow, and J. Gustafson, "Modern speech synthesis for phonetic sciences: a discussion and an evaluation," in *Proc. ICPhS*, 2019, pp. 487–491.

[36] O. Watts, G. E. Henter, J. Fong, and C. Valentini-Botinhao, "Where do the improvements come from in sequence-to-sequence neural TTS?" in *Proc. SSW*, 2019, pp. 217–222.

[37] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," https://doi.org/10.7488/ds/1994, 2017.