Speech Stimulus Continuum Generation: A Deep Learning Approach

1st Zhu Li
School of Information Science
Beijing Language and Culture University
Beijing, China
lzblcu19@gmail.com

2nd Yanlu Xie
School of Information Science
Beijing Language and Culture University
Beijing, China
xieyanlu@blcu.edu.cn

3rd Dengfeng Ke
School of Information Science
Beijing Language and Culture University
Beijing, China
dengfeng.ke@blcu.edu.cn

***Abstract*—Creating a naturalistic speech stimulus continuum (i.e., a series of stimuli equally spaced along a specific acoustic dimension between two given categories) is an indispensable component in categorical perception studies. A common method is to manually modify the key acoustic parameter of speech sounds, yet the generation process is time-consuming and the quality of synthetic speech is still unsatisfying. In this paper, we adopt an adversarial learning framework to separate the specific acoustic feature from other contents in speech signals and achieve controllable speech stimulus generation by sampling from the latent space of the key acoustic feature. Specifically, in a case study of tone continuum generation, an autoencoder was trained first to learn pitch-independent latent representations and disentangled representation of pitch separately using another auxiliary pitch predictor to regularize the latent representation. Then several latent representations between the two reference pitch representations of the continuum endpoints were sampled equidistantly. The decoder merged the pitch-independent content and a sampled latent representation to recompose an intermediate speech stimulus. Experiments on stimulus generation of tone continuum validate the effectiveness of our proposed method in both objective and subjective evaluations.**

***Index Terms*—speech synthesis, disentangled representation, adversarial training, fundamental frequency, continuum**

I. Introduction

Understanding how people represent speech categories is a central problem in speech perception research. Conducting related experiments typically requires participants to perceive speech stimulus continua (i.e., a series of stimuli equally spaced along a specific physical dimension). Synthesis of intermediate stimuli is fundamental for researchers to investigate whether and how the acoustic feature in question is utilized by listeners to identify the speech sounds. Most of the studies focusing on categorical perception have used manually modified or synthetic stimuli as experimental materials [1]–[5]. For example, in studies of Cantonese lexical tone perception [1], researchers first selected the most natural sounding syllable as a model for determining pitch-independent acoustic parameters (e.g., formant frequencies, syllable duration, and amplitude envelope of the synthetic syllables), and then in combination with these predefined parameters, a specific syllable was synthesized given an intermediate f0 contour. In studies of Mandarin Chinese tone perception [2], tone continua were constructed using naturally produced syllables for the continuum endpoints and the intermediate tone contours were obtained by interpolating key acoustic parameters (i.e., both intensity and f0 dimensions) at each time point between endpoints. Although these methods have been widely used in perceptual experiments and shown effective, it is likely that large modifications in the key acoustic parameter(s) alter the naturalness of speech stimuli, which has a considerable impact on the outcome of perceptual experiments. In addition, since acoustic parameters are continuous physical variables, performing interpolation by hand for key acoustic features may obscure subtle but important dynamic variations that listeners use as discriminative cues [6], [7].

Recently, deep probabilistic generative models have achieved impressive success. Generative Adversarial Networks (GANs) [8] and Variational Autoencoders (VAEs) [9], as powerful frameworks for deep generative model learning, have been widely applied in computer vision tasks and have provided a way to compactly represent natural image features [10]–[12]. Notably, the adversarial learning framework has been successfully used to learn disentangled image attributes in computer vision tasks [13]. The latent space of GANs for face synthesis has been shown to be able to learn disentangled and controllable facial attribute representations after linear transformations [14]. In addition, images created by sampling in deep feature spaces have been successfully used in psychological experiments to capture human category representations [15]. In voice conversion tasks, the adversarial learning approach has been effectively utilized to learn disentangled audio representations [16]. However, less work has been devoted to controllable speech stimulus generation due to the demand for high controllability of various parameters of speech signals [17], [18].

In this paper, we adopt an adversarial learning framework to separate the specific acoustic feature from other contents in speech signals. After training, our model can generate highly natural intermediate speech stimuli by merging the speech content independent of the key acoustic parameter and the sampled latent representation from the latent space of the acoustic parameter.

There are three major contributions in this paper. First, our work bridges the gap between traditional stimulus generation in speech category perception studies and state-of-the-

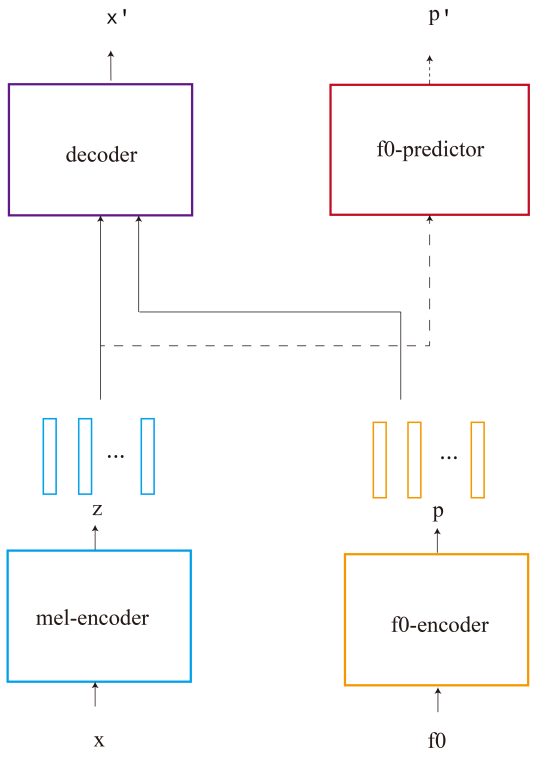

Fig. 1. The training procedure. In stage 1, the mel-encoder is trained to learn a mel representation and f0-encoder is trained to learn an f0 representation. In stage 2, the f0-predictor predict the probability that f0 is the true value. The solid line represents stage 1, the dashed line represents stage 2.

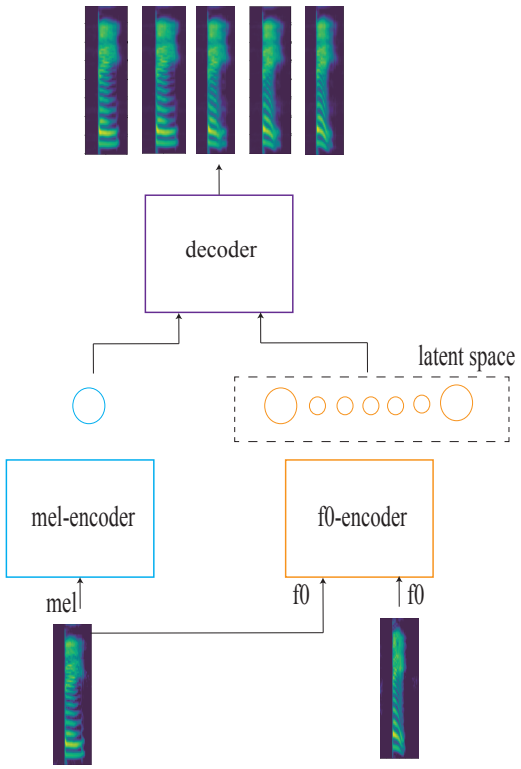

Fig. 2. Speech stimulus generation.

art generative models in deep learning (DL). Second, instead of directly performing manipulation on key acoustic cues, our proposed approach performs interpolation on latent space after disentangling the key acoustic features, which avoids possible information loss and problems of unnaturalness caused by manual interpolation. Third, we conduct one case study on tone continuum generation and the results prove the effectiveness of our method.

II. Proposed Framework

The modules for disentangling pitch from other acoustic information are adapted from the adversarial learning framework for voice conversion [16], as shown in Fig. 1. The theoretical assumption here is as follows: let $x \in \mathcal{X}$ be a sequence of multiple acoustic features where $\mathcal{X}$ is the collection of all such sequences, and $y \in \mathcal{Y}$ be a pitch sequence produced in parallel with other pitch-independent acoustic features. $\mathcal{Y}$ is the group of all pitch sequences. The training set $\mathcal{D} = \{(x_1, y_1) \dots (x_m, y_m)\}$ contains m pairs of $(x_i, y_i) \in (\mathcal{X}, \mathcal{Y})$. The modules for stimulus continuum generation are shown in Fig. 2.

A. The Training Procedure

As shown in Fig. 1, the training procedure includes learning a mel-encoder, an f0-encoder, an f0-predictor and a decoder, which contain three stages.

1) Stage 1: mel-encoder + f0-encoder + decoder: The mel-encoder is trained to map an input sequence x to a latent representation z. The f0-encoder is trained to map an f0 sequence to a latent representation p. The decoder is trained to generate x' which is a reconstruction of x from z given the specific pitch information p, as shown in (1).

$$x' = \mathrm{dec}(z, p) \tag{1}$$

The mean absolute error (MAE) between the original input sequence and the reconstruction was minimized during the training process. The reconstruction loss was computed as in (2), in which $\theta_{\text{mel-enc}}$, $\theta_{\text{f0-enc}}$ and θ_{dec} are the parameters of the mel-encoder, the f0-encoder and the decoder respectively.

$$L_{rec}\left(\theta_{mel-enc}, \theta_{f0-enc}, \theta_{dec}\right) = \sum_{(x,y)\in\mathcal{D}} |x' - x| \tag{2}$$

2) Stage 2: f0-predictor: The latent representation learned by the mel-encoder cannot be guaranteed to be pitch-independent with the guidance of (2). Given that the f0 patterns of the original sound x exist in the mel-encoder, it is likely that the intermediate stimulus generation will be unsuccessful. For this reason, we additionally trained an f0 predictor to regularize the autoencoder to make the output of the mel-encoder pitch-independent. In this stage, the parameters of the three pre-trained models in stage 1 were fixed and an f0-predictor is trained to minimize the cross-entropy loss between

the output of the f0-encoder and that of the f0-predictor based on (3).

$$L_{f0-predictor}(\theta_{mel-enc}, \theta_{f0-predictor}) = \sum_{(x_i,y_i)\in\mathcal{D}} -logP(y_i|mel - enc(x_i)) \quad (3)$$

3) Stage 3: Disentanglement: In this stage, the mel-encoder and the decoder were treated as a generator and the f0-predictor as a discriminator, which were adversarially trained to disentangle f0 information from mel representations. On the whole, the loss of the discriminator is to minimize the negative log-probability to predict f0 values, as shown in (3).

On the other hand, however, the mel-encoder is trained to maximize (3) in order to remove the pitch information in enc(x). The full objective for the autoencoder regularized by the f0-predictor is shown in (4).

$$L_{gen}(\theta_{mel-enc}, \theta_{decoder}) = L_{rec}\left(\theta_{mel-enc}, \theta_{dec}\right) - \lambda * L_{f0-predictor}(\theta_{f0-predictor}) \quad (4)$$

B. Stimulus Continuum Generation Procedure

The flow of our proposed DL-based method for synthesizing the continuum stimuli is shown in Fig. 2. First of all, the trained model in the training phase is used to give two categories of speech (x1 and x2); the pitch-unrelated representation of x1 is obtained with the mel-encoder; interpolation is done on pitch representations of the two categories (i.e., the latent space of x1 and x2 obtained with the f0-encoder). Then, mel spectrograms of the x1-x2 continuum stimuli are synthesized by sending them to the decoder together with the pitch-independent representation of x1, and finally reduced to waveforms using the vocoder, as shown in Fig. 2.

III. Experiments

A set of comparison experiments was conducted to synthesize the lexical tone continuum, using our proposed approach built upon deep learning and the traditional approach based on signal processing respectively.

A. Deep learning approach

The data for this study were based on recordings taken from the BLCU-SAIT speech corpus [19]. The corpus consists of both native and nonnative speech with monosyllabic and disyllabic words and multi-syllabic sentences. We selected the single-syllable speech data produced by a female native speaker, totaling 1520 monosyllabic words that cover all possible tones and initials in Mandarin.

For acoustic feature extraction, all audio recordings were downsampled to 16kHz, 80-dim mel spectrograms. Both mel spectrograms and f0 values were calculated with 25ms Hanning window, 6.25ms frame shift and 1024-point fast Fourier transform. The f0 values were extracted using the WORLD vocoder [20]. Pitch parameters were set at a minimum of 50 Hz and a maximum of 600 Hz in the WORLD vocoder for reliable f0 extraction.

The model was trained with the Adam optimizer and the initial learning rate was set to 0.0001. All configurations were trained for a maximum of 20000 iterations with a batch size of 48 spoken words. In the training phase, all utterances were randomly selected and pre-processed into fixed-length frames, here we set it to 64.

TABLE I
MEL-ENCODER

conv-bank block	Conv1d-bank-8, LReLU, IN
conv block × 3	C-512-5, LReLU C-512-5, stride=1, LReLU, IN, Res
dense block × 4	FC-512, IN, Res
recurrent layer	bi-directional GRU-512
combine layer	recurrent output + dense output

TABLE II
DECODER

conv block × 3	C-512-5, LReLU C-512-5, stride=1, LReLU, IN, Res
dense block × 4	FC-512, IN, Res
recurrent layer	bi-directional GRU-512
combine layer	recurrent output + dense output

TABLE III
F0-ENCODER

conv block × 4	C-512-5, LReLU C-512-5, stride=1, LReLU, IN, Res
softmax layer	$FC - N_{f0}$

TABLE IV
F0-PREDICTOR

conv block × 4	C-512-5, LReLU C-512-5, stride=1, LReLU, IN, Res
softmax layer	$FC - N_{f0}$

B. Baseline DSP approach

In this control experiment, we used the common digital signal processing (DSP) method for lexical tone continuum stimulus synthesis as the baseline. Similar to the deep learning approach, we selected /bei1/ and /bei4/ as the two reference stimuli and extracted the f0, sp, ap parameters from the two speech sounds using the WORLD vocoder, and then linearly interpolated between two sequences of extracted f0s. Finally, a new set of f0, sp, and ap obtained after interpolation was restored back to the speech signal using the WORLD vocoder.

IV. Results and Discussion

A. Objective evaluation

1) generation for different stages: As described in the previous section, the training procedure consists of three stages. Fig. 3(a) shows the first stage, which is before adversarial training. It can be seen that although the f0 information of

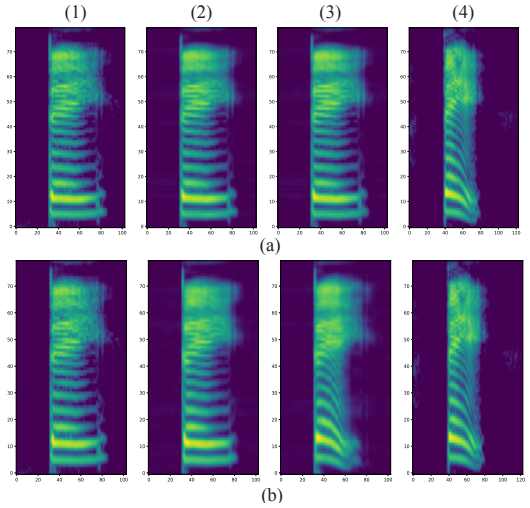

Fig. 3. The mel-spectrogram: (a) mel generation in stage1 (before adversarial training). (b) mel generation in stage3 (after adversarial training). (1) original /bei1/. (4) original /bei4/. (2) and (3) are generated mel-spectrograms: (2) is generated with the f0 information of /bei1/, (3) is generated with the f0 information of /bei4/.

/bei1/ has been replaced by that of /bei4/, the generated mel-spectrogram still contains the f0s of the original /bei1/ (see Fig. 3 a(3)). Fig. 3(b) reflects the third stage with adversarial training. It can be clearly observed that the f0 information in the generated mel-spectrogram is identical to the f0 of /bei4/ (see Fig. 3 b(3)), indicating that the pitch information has been successfully removed from the mel-spectrogram information by the disentanglement.

2) generation on latent space: Fig. 4(a) is the mel-spectrograms of continuum stimuli from /bei1/ to /bei4/ synthesized by the baseline method. Fig. 4(b) is the mel-spectrograms of continuum stimuli synthesized by the deep learning method. It can be easily observed that the samples generated using our method between /bei1/ and /bei4/ are more natural as they contain diversified distribution over all frequencies.

B. Subjective evaluation

1) Identification test: To compare our proposed approach and the traditional method of directly adjusting the acoustic parameters, an identification experiment was conducted to explore whether differences exist in categorical boundary position and width using the stimulus continua created by these two techniques. Subjects were five native speakers of Mandarin Chinese with a Mandarin level above 2A. At the beginning of the test, two reference sounds (coded as "Sound 1" and "Sound 2" respectively) were played two times to participants, and they were instructed to familiarise themselves with the two representative sounds as best as possible. The stimulus samples of each continuum were presented to the participants randomly. And participants were asked to press key "1" when they thought the sound was "Sound 1" and to press key "2" when they thought they had heard "Sound 2". The 10 stimuli were played randomly in a block. There were

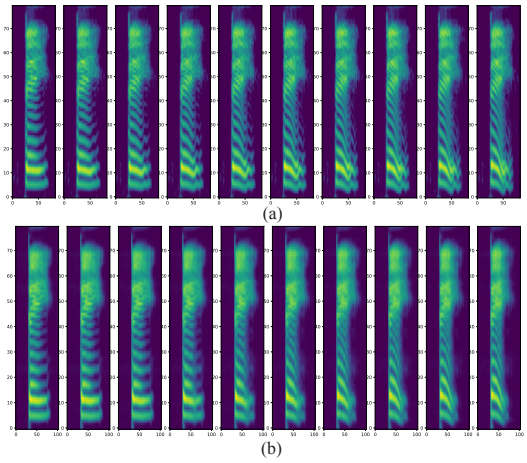

Fig. 4. (a) /bei1/ to /bei4/ tone continuum generated by the conventional method. (b) /bei1/ to /bei4/ tone continuum generated by our proposed method.

5 such testing blocks for each continuum generated by the two methods. On the whole, there is no distinct difference between the two identification patterns, suggesting that both approaches can generate relatively natural stimulus samples with the acceptable sound quality for perception studies.

2) MOS evaluation: The overall quality of the stimulus samples generated by the two methods was evaluated using the mean opinion score (MOS). Five native speakers of Mandarin Chinese were recruited, and none of them had participated in the previous experiment. Listeners were asked to rate the overall naturalness of the stimulus samples on a scale from 1 and 5. A total of 80 voice stimuli (8 continua) were mixed and presented to listeners in a random order. Based on the identification results, generated sounds were divided into within-category stimuli and between-category stimuli. For the tone continuum, the third, fourth and fifth stimuli were viewed as between-category. In accordance with this classification, the overall MOS, within-category MOS and between-category MOS were calculated. Tabel V summarizes the results of listeners' evaluation of the synthesized stimuli.

TABLE V
MOS OF STIMULUS SAMPLES

System	MOS (overall)	between category	within category
/bei1/-/bei4/ (Baseline)	3.67	4.15	3.28
/bei1/-/bei4/ (Deep learning)	**3.88**	**4.27**	**3.44**

Pairwise comparisons using the paired Mann-Whitney U Tests [21] show that between-group differences in overall MOS are not significant ($p > 0.05$). This suggests that generally speaking, the quality of generated stimuli using the proposed approach is comparable to that of the DSP approach. Notably, the between-category MOS of the tone1-tone4 continuum based on the deep learning approach are slightly higher. These results indicate that both approaches can generate relatively natural stimulus samples with acceptable sound quality, but for those stimuli near the category boundary,

the deep learning method slightly improves the naturalness of generated speech over the DSP baseline. [1]

V. Conclusion

In this paper, we adopted a deep learning approach to produce stimulus continua based on an adversarial learning framework. This work bridges stimulus continuum generation and state-of-the-art DL techniques. We conducted case studies on stimulus synthesis of tone continua using our proposed model. The results indicated that our approach yielded more natural between-category stimuli compared to manual manipulation on the key acoustic feature. Future directions include disentangling more fine-grained acoustic features (e.g., formants for generation of vowel continua).

Acknowledgment

This work is supported by the Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities)(16ZDJ03), the Fundamental Research Funds for the Central Universities, and the Research Funds of Beijing Language and Culture University (21YCX167), National Social Science Foundation of China (18BYY124), Discipline Team Support Program of Beijing Language and Culture University (GF201906), and Wutong Innovation Platform of Beijing Language and Culture University (19PT04).

References

[1] A. L. Francis, V. Ciocca, and B. K. C. Ng, "On the (non) categorical perception of lexical tones," *Perception & psychophysics*, vol. 65, no. 7, pp. 1029–1044, 2003.

[2] P. A. Hallé, Y.-C. Chang, and C. T. Best, "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners," *Journal of phonetics*, vol. 32, no. 3, pp. 395–421, 2004.

[3] W. Strange and S. Dittmann, "Effects of discrimination training on the perception of/rl/by Japanese adults learning English," *Perception & psychophysics*, vol. 36, no. 2, pp. 131–145, 1984.

[4] K. S. MacKain, C. T. Best, and W. Strange, "Categorical perception of English/r/and/l/by Japanese bilinguals," *Applied psycholinguistics*, vol. 2, no. 4, pp. 369–390, 1981.

[5] V. G. Skuk and S. R. Schweinberger, "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," 2014.

[6] M. Gubian, F. Cangemi, and L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," 2010.

[7] M. Gubian, Y. Asano, S. Asaridou, and F. Cangemi, "Rapid and smooth pitch contour manipulation," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013, pp. 31–35.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[10] T. White, "Sampling generative networks," *arXiv preprint arXiv:1609.04468*, 2016.

[11] T. Sainburg, M. Thielk, B. Theilman, B. Migliori, and T. Gentner, "Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions," *arXiv preprint arXiv:1807.06650*, 2018.

[12] Y.-C. Chen, X. Xu, Z. Tian, and J. Jia, "Homomorphic latent space interpolation for unpaired image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2408–2416.

[1] Demo webpage: https://abel1802.github.io/Stimulus-with-DL/

[13] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader networks: Manipulating images by sliding attributes," *arXiv preprint arXiv:1706.00409*, 2017.

[14] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9243–9252.

[15] J. C. Peterson, J. W. Suchow, K. Aghi, A. Y. Ku, and T. L. Griffiths, "Capturing human category representations by sampling in deep feature spaces," *arXiv preprint arXiv:1805.07644*, 2018.

[16] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[17] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[18] M. Tomczak, M. Goto, and J. Hockman, "Drum synthesis and rhythmic transformation with adversarial autoencoders," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2427–2435.

[19] B. Wu, Y. Xie, L. Lu, C. Cao, and J. Zhang, "The construction of a Chinese interlanguage corpus," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 183–187.

[20] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[21] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores." in *Interspeech*, 2017, pp. 3976–3980.