



Data requirements, selection and augmentation for DNN-based speech synthesis from crowdsourced data

Markus Toman¹, Geoffrey S. Meltzner¹, Rupal Patel¹

¹VocaliD, Inc.

markus@vocalid.co, geoff@vocalid.co, rupal@vocalid.co

Abstract

Crowdsourcing speech recordings provides unique opportunities and challenges for personalized speech synthesis as it allows gathering of large quantities of data but with a huge variety in quality. Manual methods for data selection and cleaning quickly become infeasible, especially when producing larger quantities of voices. We present and analyze approaches for data selection and augmentation to cope with this. For differently-sized training sets, we assess speaker adaptation by transfer learning, including layer freezing, and sentence selection using maximum likelihood of forced alignment. The methodological framework utilizes statistical parametric speech synthesis based on Deep Neural Networks (DNNs). We compare objective scores for 576 voice models, representing all condition combinations. For a constrained set of conditions we also present results from a subjective listening test. We show that speaker adaptation improves overall quality in nearly all cases, sentence selection helps detecting recording errors, and layer freezing proves to be ineffective in our system. We also found that while Mel-Cepstral Distortion (MCD) does not correlate with listener preference across the range of values, the most preferred voices also exhibited the lowest values for MCD. These findings have implications on scalable methods of customized voice building and clinical applications with sparse data.

Index Terms: speech synthesis, noisy data, data selection

1. Introduction

Typical speech corpus collection is a time and resource intensive process that involves recording a small number of speakers (often professional voice talent) over the course of several sessions, using professional recording equipment in sound recording studios. The resulting corpus is high in audio quality but limited in vocal diversity. However, with the proliferation of inexpensive but high-quality recording technology combined with the advent of cloud computing, crowdsourcing speech recording can yield sizeable corpora. Crowdsourcing also allows to capture language varieties such as dialects, foreign accents and sociolects which are all important aspects of the personality of a voice. However, along with the benefits of crowdsourced speech comes its own challenges. Crowdsourced corpora are susceptible to poor audio quality as recordings are essentially conducted in the wild - by novice voice talent using consumer-grade equipment for recording in their homes, offices or schools. Furthermore, because contributions are often conducted over the course of several sessions, the recording environments can vary from session to session. The unsupervised manner of data collection also means that discrepancies between the expected prompt and the actual spoken utterance cannot be corrected during the recording session (i.e. by repeating the proper content) which leads to errors in the synthesis. Finally the crowdsourced collection relies on the willingness

and availability of speakers to complete the recordings. Although algorithmic methods can be and have been developed to mitigate some of these challenges, reducing the size of the training corpus can potentially address many of the limitations cited above. A small enough corpus allows speakers to complete the corpus in a single session, reduces the probability of transcription mismatch, and enables small scale quality monitoring to be useful. These properties are particularly important when producing larger quantities of voices. However, using small corpora limits the amount of available training data and necessitates the use of different synthesis methods (e.g. adaptive methods). Furthermore, these adaptive voices need to be assessed to ensure that their quality is equivalent to voices generated by training on much larger corpora. These requirements thus motivated the current study in which we assessed, both objectively and perceptually, speech data selection and training methods. This paper is organized as follows. We first describe the Text-to-Speech (TTS) system used to generate the experimental speech data, followed by the data selection and speaker adaptation processes. We then describe the objective and perceptual evaluations of the resulting synthetic speech, followed by results and conclusion.

2. System description

The system used for the conducted experiments is subject to computational constraints as it has to synthesize speech faster than realtime on older mobile devices common with VocaliD customers. We used the Merlin framework [1] for building voice models and the WORLD vocoder [2] for parameterization. Synthesis was conducted using the VocaliD TTS engine. Reproducibility is still given, as the output of the engine is indistinguishable from the output of Merlin. The basic ("demo") recipe from Merlin with a few modifications was used for all experiments: We used a much smaller network of 6 layers with 128 units each. Informal subjective listening tests indicated that this minimally impacted quality while making the training procedure less prone to overfitting. The reduced number of parameters (a reduction from 5,875,899 to 161,211) sped up training, improved time-to-speak in the live system and also reduced the model size from about 80 Mb to 20 Mb. The reduced training time was essential to produce the 576 voice models in reasonable time. Furthermore, we used a separate duration model of 6 layers with 1024 units, *tanh* activation throughout, a learning rate of 0.002, mini-batch size 256. Training was conducted for 35 epochs or until the error on the validation set stopped decreasing. The first 10 epochs were warmup epochs with a momentum of 0.6, then momentum of 0.9 was used with an exponentially decaying learning rate. For forced alignment HTK [3] was used, we modified the standard Merlin recipe to allow the use of larger data sets and enable parallel processing by splitting the data set. All optimizations implemented for the

experiments were merged back into the original Merlin repository. Alignment was conducted only once per speaker, so the set of labels was fixed for all experiments. This was done to avoid introducing yet another variable factor - i.e. the results of the voice training shall not be influenced by the performance of the alignment. Similarly, a fixed test set of 300 sentences was used for all experiments.

We further employed a simple sentence recombination method because users tend to synthesize longer sentences than those contained in the training set. In the well-known CMU ARCTIC corpus¹, the longest sentence contains 15 words, with a median of 9 words. This is by design, because longer sentences increase difficulty for the speakers, resulting in more errors. When synthesizing novel, longer sentences, min-max normalization clips quantitative linguistic features (e.g. "number of sentences/words in the current phrase") too early, resulting in unnatural prosody. Therefore we augment the training data by randomly selecting approximately 5% of the sentences from the training set and concatenating them. On the waveform we apply silence pruning using sox², append a 50ms silence to the first waveform and then concatenate the second waveform. In the respective orthographic transcription of the first sentence we replace the final punctuation with a comma and then append the transcription of the second sentence. This procedure was repeated twice. Informal listening showed improvements on long sentences, so we enabled this mechanism in the system assessed here.

3. Methods

3.1. Data selection

When dealing with crowdsourced data, we assume that a significant part of the recordings is of uncertain quality. This not only encompasses acoustic aspects like background noise and reverberation but also mispronunciations, truncated sentences, and meaningless babbling. Here, we focus on discrepancies between expected prompts and the actual spoken utterance. We conducted a forced alignment using Hidden-Markov Models (HMMs) to align phone state sequences (5 states per phone) with the waveforms. We used the Viterbi algorithm via the HTK tool "HVite" to find the most likely sequence of state durations for each sentence with respect to the sequence of input phones, the recordings, and an HMM previously estimated from the data. For each state sequence we consider the average logarithmic likelihood (of the states emitting the given observations) as predictor for sentence quality.

As an initial experiment to assess the general feasibility of this approach, we injected 10 sentences into the training set, for which the orthography did not match the waveform. Typical cases where this might occur are errors in the recording procedure or malicious speakers. We selected a speaker with high variance of recording quality (speaker *F1* in Section 4). We found that the 4 sentences with the lowest score were from the injected set, and all injected sentences were within the worst 48 (of 4316) sentences. This corresponds to the 0.012 low quantile and suggests a detection accuracy justifying a more exhaustive experiment. For the experiments we aligned the the full data set and then selected the n sentences with the highest likelihood. For each n , a voice model was trained. The hypothesis was that the likelihood would be a useful predictor for certain aspects of quality.

¹CMU ARCTIC: http://www.festvox.org/cmu_arctic/

²sox: <http://sox.sourceforge.net/>

3.2. Data augmentation

A well-known method for data augmentation is speaker adaptation, where the most common approach is to build an average voice model of multiple speakers and then adapt a model for new (target) speaker from it. Speaker adaptation is a well-researched topic in HMM-based speech synthesis [4, 5, 6, 7, 8, 9] but still relatively unexplored for DNN-based synthesis. Arik et al [10] found that speaker adaptation by fine-tuning (i.e. transfer learning), while most demanding with regards to computational resources, gives the best results in terms of naturalness. In [11], a multi-speaker model is used where each speaker uses a separate output layer, fine-tuning is then used with adding a new output layer for the adapted speaker. Here we investigate transfer learning by training a single DNN with data from multiple speakers, then fine-tuning the network weights with data from the target speaker. We also assess the effects of layer freezing, i.e. keeping some layer weights fixed during the training.

In our experiments, an average model trained from about 25,000 sentences from 6 speakers of different ages and genders was used. The speakers were selected by manually listening to recordings from the crowdsourced data set. Alignment of the average voice model was conducted using per-speaker normalization. The network weights of this average model were then used as initial weights when training the target speaker model

4. Experiments

We selected 6 speakers from the VocaliD voicebank to cover different age groups and both male and female speakers. Table 1 the logarithmic likelihood means from forced alignment as well as the lowest Mel-Cepstral Distortion (MCDs) of the respective voice models produced during the experiments.

Table 1: *Speakers used in the experiments*

ID	Gender	Age	μ likelihood	min MCD
F1	female	21	-46.35	8.71
F2	female	50	-38.28	6.26
M1	male	34	-44.16	6.82
M2	male	73	-41.35	5.97
C1	female	8	-48.68	9.59
C2	male	13	-46.68	9.29

4.1. Objective evaluation

We calculated the MCD on the 300 sentence test set for all generated voices. MCD is commonly used to assess synthetic speech and has been shown to correlate with perceived voice quality [12, 13]. We use the implementation as enabled by default in Merlin. We assessed the effect of the sentence selection strategies ("best", i.e. the best sentences according to the alignment likelihoods, and "random") and of speaker adaptation ("non-adapted" vs. "adapted"). We trained voices for all 6 speakers in all 4 combinations for the following selected sentence counts in the training set: (50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000), with 4000 corresponding to the full data set. So for each speaker this resulted in $12 * 2 * 2 = 48$ voices, $48 * 6 = 288$ in total.

In Figure 1 the results for all 6 speakers are presented. We found that "adapted" was superior to "non-adapted" in general, especially for smaller training sets. As expected, the differences became smaller as more data was used. The combination of

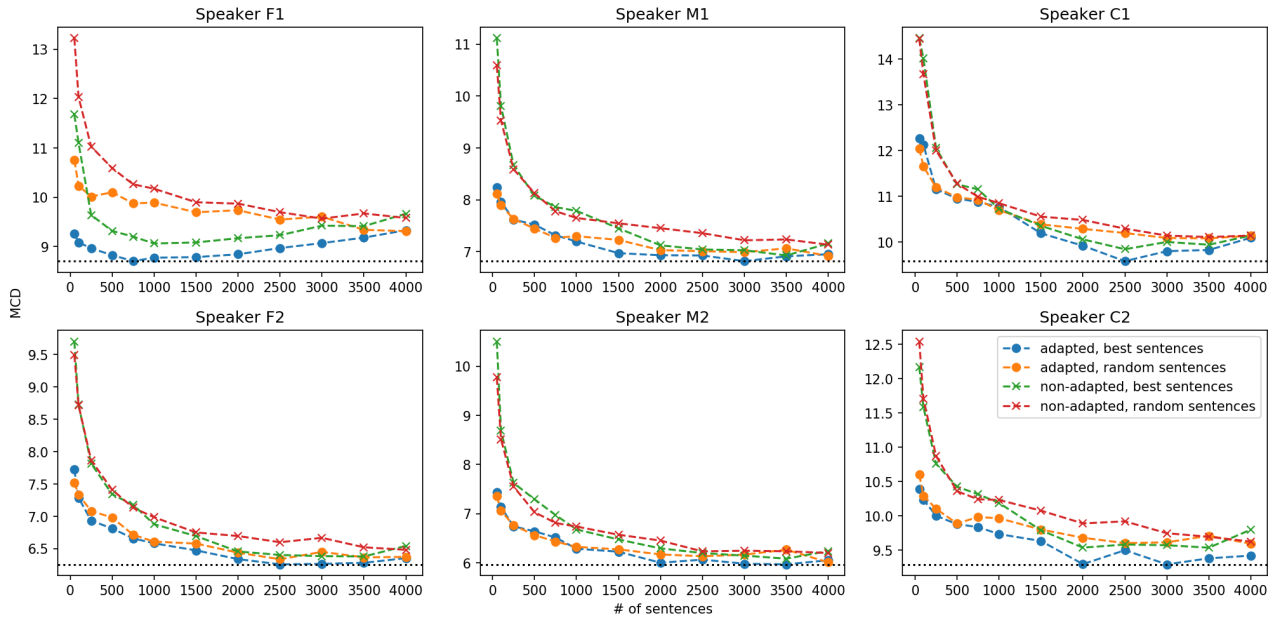


Figure 1: Objective results for all speakers, showing number of sentences in training set and MCD

”adapted” with ”best” generally performed at least as well as all other conditions. While for speakers with consistent quality we rarely found a significant difference in MCD when using the ”best” strategy, the difference was most remarkable for speaker *F1*. As expected, when more sentences with lower likelihood were added to the training set, the sentence selection strategies began to converge. Speakers *C1* and *C2* were children with a higher variance in recording quality, which is reflected in Figure 1, showing much less stable graphs. Generally, the results Table 1 suggest that higher likelihood means in alignment yielded voices with lower MCDs.

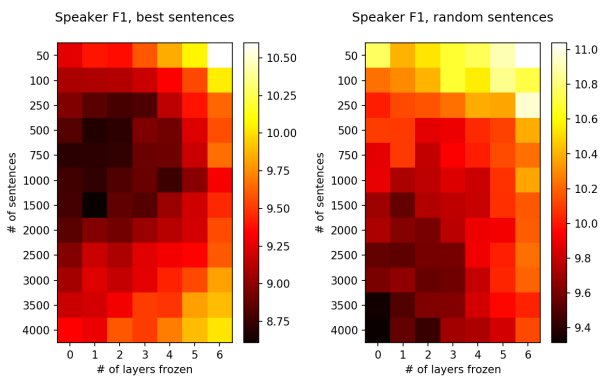


Figure 2: Speaker adaptation with layer freezing, speaker *F1*

We further examined the effect of layer freezing on synthetic speech quality. As this greatly increases the number of possible combinations, we selected speakers *F1* and *M1* for this analysis. Freezing 0 to 6 layers results in 7 possible conditions. With the other conditions we end up with $7 * 2 * 12 * 2 = 336$ voice models. The resulting MCDs are shown in Figures 2 and 3 for speakers *F1* and *M1* respectively. Column 0 in both cases corresponds to the ”adapt” graphs shown in Figure 1. It is clear that the lowest MCD scores concentrate around 0 or 1

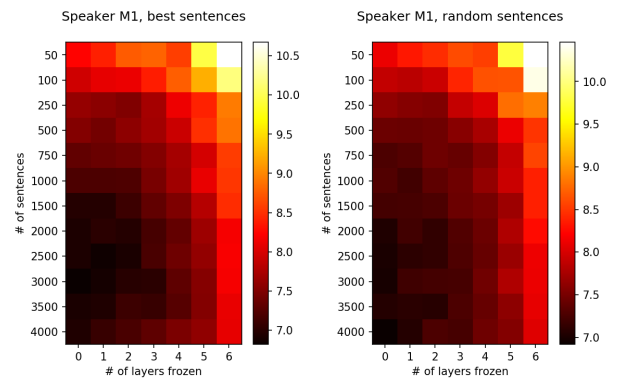


Figure 3: Speaker adaptation with layer freezing, speaker *M1*

frozen layers. Again, the optimal number of ”best” sentences in the training set varied greatly between speakers *F1* and *M1*, with the lowest MCD for speaker *F1* being at 1 frozen layer, 1500 sentences and ”best” sentence strategy; the lowest MCD for speaker *M1* being at 300 sentences, 0 frozen layers, 3000 sentences. While two speakers are not sufficient for drawing conclusions, these initial results suggest that freezing more than 1 layer generally seemed to increase rather than decrease error.

4.2. Perceptual evaluation

To assess the efficacy of using the MCD as an objective measure of speech naturalness, we conducted a set of perceptual experiments on speech tokens generated from synthetic voices trained on 5 different sentence counts for each speaker using the Method of Paired Comparisons (PC). Specifically, for each speaker, we generated tokens from the Harvard Sentences List (Group 72) [14], for voices trained using different sized corpora for the adapted and non-adapted training methods and the ”best” and ”random” sentence selection methods, resulting in

4 conditions for each speaker. Pairwise combinations of tokens were presented to a total of 46 listeners (all of whom were native American English speakers with no reported history of hearing loss) using Amazon Mechanical Turk³ such that 25 judgments were made for each of the 4 conditions for each of the 6 speakers (note that not every listener rated all pair combinations; 600 is the total number of judgments made). For each presented pair, the listeners were instructed to select the speech token that sounded "most natural." Both the order of the pairs as well as the order within pairs were randomized, and 10% of the pairs were repeated to assess listener reliability. The data collected from the PC procedure were analyzed using Thurstone's Law of Comparative Judgment [15]. This allows for transformation of the forced choice selections into a perceptual distance scale, that not only ranks the tokens but also provides the perceptual distance between them. The standard formulation of the Law of Comparative Judgment requires that no one token type is judged to always be "better" or "worse" than all the others. In this study, this necessary requirement was violated a number of times, forcing the use of an alternative analysis method [16] which has been successfully used in other speech perception studies [17].

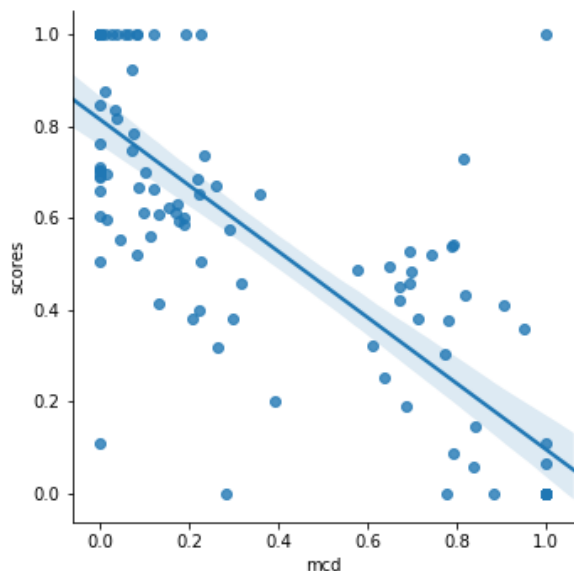


Figure 4: Pairwise scores from subjective evaluation vs. MCD, both normalized within each condition, for all speakers

Figure 4 shows the results from the subjective evaluation plotted against the results from the objective MCD-based evaluation. To be able to aggregate the data from all speakers, MCD and preference scores have been min-max normalized (within each condition) to 0.0 - 1.0 range. Thus, a subjective score of 1.0 represents a voice model that was preferred most over all other models in the same condition. An MCD of 1.0 represents the highest MCD value for a given combination of conditions (i.e. the highest error). A cluster at subjective score 1.0 and MCD near to 0.0 shows that the voices most preferred by listeners also exhibit low values of MCD. More precisely, the mean normalized MCD of all voices most preferred by the listeners is 0.079 while the mean MCD of the least preferred voices is 0.96.

³Amazon Mechanical Turk: <https://www.mturk.com/>

Most variance in the data was caused by speaker $F1$, who was also responsible for the extreme outlier at $x = y = 1.0$ corresponding to the 4000 sentences voice with "best" and "adapted" conditions. So while the overall coefficient of determination r^2 is -0.81 , r^2 for speaker $F1$ is only -0.52 . The other values for r^2 are $F2 : -0.90$, $M1 : -0.77$, $M2 : -0.91$, $C1 : -0.87$, $C2 : -0.89$.

We found overall intra-subject reliability to be 65%, with most of the lack of reliability being generated by evaluating adapted pairs (55%). Reliability for non-adapted pairs was much higher at 85%. These scores are consistent with the respective MCD curves for adapted and non-adapted tokens shown in Figure 1. The range of MCD scores for adapted tokens was notably smaller than that of non-adapted tokens, suggesting that they were of similar acoustic quality, making it more difficult to consistently differentiate between them.

5. Discussion and Conclusion

Our investigation and assessment of several strategies of data selection and augmentation strategies have led to important insights that can guide the synthesis of voices from crowdsourced speech data, or more generally, speech data collected in the wild. Our objective evaluation of sentence selection by identifying the best sentences according to forced alignment likelihood suggests that DNNs are very robust to noise, as long as enough data of reasonable quality is available. Adding low-quality data only resulted in a higher MCD for speakers with generally noisy data. Furthermore, speaker adaptation by fine-tuning an existing average model generally had a much stronger effect on the MCD than did data selection. Our perceptual evaluations demonstrated that MCD is a reasonable predictor for listener preference; while there is some variance, the voice models most preferred by listeners also exhibited the lowest MCD values.

Overall, these results suggest that optimizing data selection can be beneficial when the data quality is expected to be very low. Without data selection, it seems important to use adaptation with all available data by default when data is sparse (i.e. up to 4000 sentences). Furthermore, for our 6 speakers we saw that higher alignment likelihoods resulted in lower MCDs of the final voices. As our results suggest that lower MCD scores predict better quality voices, likelihood scores could be employed for training large average models where managing training time and model capacity becomes relevant (e.g. selecting the 500,000 best sentences from a 5,000,000 sentence database). They can also be used to identify promising speakers from large crowdsourced databases more efficiently.

Demonstrating that MCDs predict synthetic speech quality establishes an experimental paradigm for rapidly evaluating the naturalness of synthetic voices. This can be particularly useful when testing a large number of voicebuilding parameters to establish an optimal configuration and also for automating quality control when building a large number of synthetic voices. Future work will focus on extending these results to a large set of speakers and establishing how well MCD predicts synthetic speech intelligibility.

6. Acknowledgment

The authors would like to thank Steven Rife for designing and conducting the perceptual experiments on Amazon Mechanical Turk. This work was supported in part by NIDCD grant R44 DC014607-02.

7. References

- [1] Z. Wu, O. Watts, and S. King, *Merlin: An Open Source Neural Network Speech Synthesis System*, 9 2016, pp. 218–223.
- [2] 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, pp. 1877–1884, 2016.
- [3] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK version 3.4)*. Cambridge, UK: Cambridge University Engineering Department, 2006.
- [4] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.
- [5] J. Yamagishi, “Average-voice-based speech synthesis,” Ph.D. dissertation, Tokyo Institute of Technology, Tokyo, Japan, 2006.
- [6] M. Wester and R. Karhila, “Speaker similarity evaluation of foreign-accented speech synthesis using hmm-based speaker adaptation,” in *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5372–5375.
- [7] R. Dall, C. Veaux, J. Yamagishi, and S. King, “Analysis of speaker clustering strategies for HMM-based speech synthesis,” in *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2012, pp. 995–998.
- [8] M. Toman and M. Pucher, “Structural KLD for cross-variety speaker adaptation in HMM-based speech synthesis,” in *Proceedings of the 10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications (SPPRA)*, Innsbruck, Austria, 2013, pp. 382–387.
- [9] M. Toman, M. Pucher, and D. Schabus, “Multi-variety adaptive acoustic modeling in HSMM-based speech synthesis,” in *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW)*, Barcelona, Spain, Aug. 2013, pp. 83–87.
- [10] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural Voice Cloning with a Few Samples,” *ArXiv e-prints*, Feb. 2018.
- [11] S. Pascual and A. Bonafonte, “Multi-output rnn-1stm for multiple speaker speech synthesis and adaptation,” in *2016 24th European Signal Processing Conference (EUSIPCO)*, Aug 2016, pp. 2325–2329.
- [12] R. F. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1. IEEE, May 1993, pp. 125–128 vol.1.
- [13] J. Kominek, T. Schultz, and A. W. Black, “Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion,” in *SLTU*, 2008.
- [14] “IEEE recommended practice for speech quality measurements,” *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, September 1969.
- [15] L. L. Thurstone, “A law of comparative judgement,” *Psychological Review*, vol. 34, pp. 273–286, 1927.
- [16] D. J. Krus and P. H. Krus, “Normal scaling of the unidimensional dominance matrices: the domain referenced model,” *Educational and Psychological Measurement*, vol. 37, no. 1, pp. 189–193, 1977.
- [17] G. S. Meltzner and R. E. Hillman, “Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech,” *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 4, pp. 766–779, 2005.