INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Training Multi-Speaker Neural Text-to-Speech Systems using Speaker-Imbalanced Speech Corpora

Hieu-Thi Luong[1,2], *Xin Wang*[1], *Junichi Yamagishi*[1,2,4], *Nobuyuki Nishizawa*[3]
[1]National Institute of Informatics, Tokyo, Japan
[2]SOKENDAI (The Graduate University for Advanced Studies), Kanagawa, Japan
[3]KDDI Research Inc., Saitama, Japan
[4]University of Edinburgh, Edinburgh, UK
{luonghieuthi,wangxin,jyamagis}@nii.ac.jp, no-nishizawa@kddi-research.jp

Abstract

When the available data of a target speaker is insufficient to train a high quality speaker-dependent neural text-to-speech (TTS) system, we can combine data from multiple speakers and train a multi-speaker TTS model instead. Many studies have shown that neural multi-speaker TTS model trained with a small amount data from multiple speakers combined can generate synthetic speech with better quality and stability than a speaker-dependent one. However when the amount of data from each speaker is highly unbalanced, the best approach to make use of the excessive data remains unknown. Our experiments showed that simply combining all available data from every speaker to train a multi-speaker model produces better than or at least similar performance to its speaker-dependent counterpart. Moreover by using an ensemble multi-speaker model, in which each subsystem is trained on a subset of available data, we can further improve the quality of the synthetic speech especially for underrepresented speakers whose training data is limited.
Index Terms: speech synthesis, multi-speaker modeling, imbalanced corpus, ensemble learning

1. Introduction

Recent advances in statistical parametric speech synthesis research have produced synthetic speech indistinguishable from natural speech when a model is trained with a large and high quality speech corpus [1, 2]. However to scale the technology to multiple voices and reduce the production cost, the ability to build TTS systems from a smaller and less refined corpus is crucial. As data sparsity is the major challenge for this task, many schemes have been proposed to alleviate it. If the speech corpus is created from scratch, the sentence corpus used for recording could be carefully designed to ensure a balanced coverage of linguistic units [3, 4]. A less refined speech corpus, such as a corpus of found data, can also be used by filtering out utterances deemed unfit [5, 6]. A data selection scheme can also be applied on legacy corpora to remove redundant samples [7, 8]. In another approach, we could combine speech data from many speakers and train a multi-speaker TTS system [9].

Recent neural acoustic models are capable of achieving high performance for both single speaker modeling [2] and multi-speaker modeling [10, 11] tasks. The multi-speaker model is simple to set up [12, 13] and can generate more stable speech waveforms than those of the speaker-dependent model when the amount of the target speaker's data is limited [10]. Latorre et al. [14] compared the performances of multi-speaker and single-speaker models using different amounts of data and reported similar results for various conditions. In these multi-speaker experiments [11, 14], the number of utterances contributed by each speaker is kept perfectly or roughly balanced. In this paper, we are interested in finding the best strategy to train a multi-speaker model using an existing speaker-unbalanced corpus.

Class imbalance is a common issue faced by many classification systems because real-world data are usually predominated by the normal classes while lacking samples of the abnormal classes. Many techniques have been proposed to tackle this problem. Over-sampling and under-sampling are simple and effective approaches to obtain a synthetically balanced corpus [15]. In this paper, we use the same techniques to prepare the training set for a multi-speaker acoustic model. Moreover, we propose using an ensemble model, which combines predictions of multiple subsystems, to produce a better prediction itself. Our ensemble acoustic model for speech synthesis shares the same spirit as the ensemble deep learning system for speech recognition [16].

In section 2 of this paper, we describe our methodology for multi-speaker acoustic and the ensemble models. Section 3 provides details about the experimental conditions and Section 4 presents both objective and subjective evaluation results of our proposal. We conclude in Section 5 with a brief summary and mention of future work.

2. Multi-speaker and ensemble models

2.1. Multi-speaker model for speaker-imbalanced corpus

In this paper we adopt the same auto-regressive neural-network acoustic model used in our prior publication [1]. By appending a one-hot vector speaker code to every frame of the linguistic input $\boldsymbol{x}$, we created a multi-speaker model that can generate multiple voices simply by changing the speaker code. The method is simple but effective and does not depend on the network architecture [13, 17]. This essentially means that all parameters of the network are shared among all training speakers except the bias of the first hidden layer:

$$\boldsymbol{h}_1 = \tanh(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{c}_1 + \boldsymbol{b}^{(k)}) \quad (1)$$

where $\boldsymbol{h}_1$ is the output of the first hidden layer containing m units, $\boldsymbol{W}_1 \in \mathbb{R}^{m\times m}$ and $\boldsymbol{c}_1 \in \mathbb{R}^{m\times 1}$ are common parameters shared among all speakers, and $\boldsymbol{b}^{(k)} \in \mathbb{R}^{m\times 1}$ is a speaker-specific bias projected from the speaker's one-hot vector. $\tanh$ is the non-linear activation function of the first hidden layer.

As most of the network parameters are shared and stochastically trained with combined data, using an imbalanced corpus might produce a model that is over-trained on the majority speakers while under-trained on the minority. To test this hypothesis we apply resampling techniques, which are widely used to create synthetically balanced datasets [18, 15]. Here,

http://dx.doi.org/10.21437/Interspeech.2019-1311

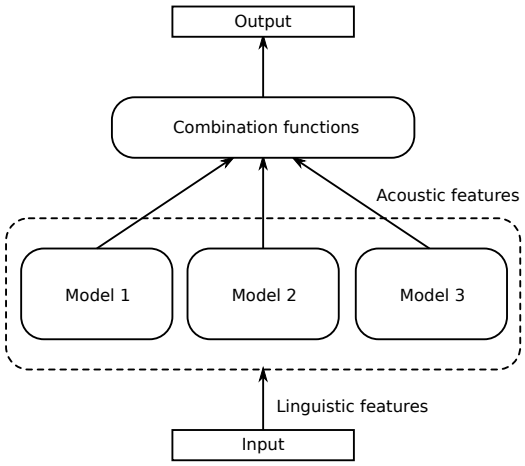

Figure 1: *Ensemble multi-speaker acoustic model used for our investigation.*

we can choose to perform under-sampling [18] of the majority speakers, over-sampling of the minority speakers [19], or a little of both [15]. While these techniques are commonly used for classification tasks, we applied them in the context of training a multi-speaker neural acoustic model.

2.2. Linear ensemble for acoustic feature inference

In addition to the resampling techniques, we also investigate using stacking [20, 21] to combine the predictions of several systems in the hope of further reducing the mismatch between generated and real-life samples. Ensemble learning is a method of using multiple models to obtain a better performance; it is used in many other research fields [22]. For example, Deng and Platt [16] performed a linear combination of the original speech-class posterior probabilities provided by subsystems at the frame level for automatic speech recognition (ASR). Their ensemble model capitalizes on the diversity of neural network architectures to provide diverse prediction outputs.

Our ensemble model, shown in Fig.1, shares many traits with the model proposed in [16] for ASR. To create diverse subsystems, we used the same network architecture in each subsystem but trained them on different data subsets randomly sampled from a training corpus. This strategy is more straightforward than creating subsystems with varied network architectures [16, 23]. Moreover we take a much simpler and non-parametric approach for the combination functions to test our hypothesis. Deterministic average-based combination functions are defined to combine the output of the subsystems. As the two main acoustic features used in our experiments are mel-generalized cepstral coefficients (MGCs) and fundamental frequency (F0), we define the combination functions as follows:

- **Combination function for MGC**: As the MGCs at each frame are continuous values, our ensemble model simply computes the average of the MGCs produced by the subsystems.
- **Combination function for F0**: Because the F0 is a continuous value at a voiced frame but a discrete symbol (i.e., unvoiced flag) at an unvoiced frame, we first decide whether one frame is voiced or unvoiced by voting. If most of the subsystems generated voiced F0s values, we take the average F0 value as the ensemble model's output. Otherwise, the output F0 is set to unvoiced.

3. Experiments

3.1. Dataset and features

Our experiments are data-driven and we seek to identify the best approach to train a speech synthesis system from an imbalanced speech corpus. The corpus we used contained utterances from ten female Japanese speakers, who are professional or at least familiar with voice acting work. The number of utterances of each speaker ranged from 1,000 to 10,000. After processing and removing utterances unsuitable for speech synthesis, we split the remaining data into training, validation and testing sets, as displayed in Table 1. As we applied a sampling technique to create a synthetic speaker-balanced corpus, the number of unique utterances of each speaker obtained from these sampling sessions are also included in Table 1.

The acoustic features used in our experiments consist of 60-dimensional Mel-generalized cepstral coefficients (MGC) and 511-bin quantized mel scale fundamental frequency (F0) plus one bin for the unvoiced case. These features are extracted from 48-kHz speech waveform using a 25-ms window and shifting 5 ms each frame. Linguistic features consist of typical Japanese linguistic information such as phonemes, moras (syllabic unit), part-of-speech tags, interrogative intention, and pitch-accent. The final linguistic features are encoded as a 265-dimensional vector for each frame including duration information extracted from forced-alignment with the acoustic feature sequence, which is obtained using an external system.

3.2. Model configurations

We adopted the same architecture described in our previous publication [1] for the acoustic models. A shallow autoregressive network (SAR) [24] is used to model MGC and a deep autoregressive network (DAR) [25] is used for quantized mel scale F0. The SAR contains two 512-unit non-linear feedforward layers followed by two 256-unit bi-directional layers, and a linear output layer. Similarly, the DAR contains two 512-unit feedforward layers, a 256-unit bi-directional recurrent layer and a 128-unit uni-directional recurrent layer that receives a feedback link from the previously generated samples and a linear layer that maps to the desired output. For the multi-speaker model, a 10-dimensional one-hot vector representing speakers is appended to every frame of the linguistic sequence. The acoustic model is trained using stochastic gradient with the utterance order shuffled to make sure the model learns the optimal representation for all speakers.

A speaker-independent WaveNet vocoder [26] was trained using the combined training data of all speakers. This model contained 40 dilated layers similar to the original WaveNet [27]. It was directly trained using the natural MGC and quantized mel-scale F0s from all the speakers, without speaker one-hot vectors. The target waveform had a sampling rate of 16 kHz and was quantized using the 10-bit μ-law standard.

3.3. Strategies for handling imbalanced corpus

The main investigation in this paper is which methodology efficiently uses an imbalanced multi-speaker corpus to improve performance for the generated speech of all speakers involved. Multiple strategies are compared in the experiments:

- **`SD`**: The conventional speaker-dependent models, each of which is trained using one target speaker's data listed in Table 1. This is our baseline strategy.
- **`UN`**: A multi-speaker model trained with an under-

Table 1: *Data sets of target speakers.*

Speaker ID	**XS01**	**XS02**	**S03**	**S04**	**S05**	**M06**	**M07**	**M08**	**L09**	**XL10**
Training (unique utterances):										
Speaker-Dependent	735	994	1393	1568	1749	3024	3983	4364	5516	8750
Sampling 1st	728	938	1227	1341	1444	1901	2088	2179	2320	2532
Sampling 2nd	729	955	1214	1340	1442	1892	2074	2185	2312	2516
Sampling 3rd	722	944	1242	1329	1418	1916	2122	2186	2325	2554
Ensemble (Sampling 1+2+3)	735	994	1391	1559	1742	2869	3541	3807	4424	5630
Validation	50	50	50	50	50	50	50	50	50	50
Testing	100	100	100	100	100	100	100	100	100	100

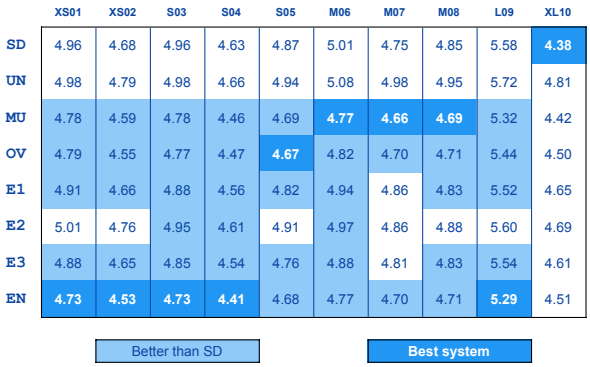

Figure 2: *Mel-ceptral distortion (smaller is better).*

Figure 3: *F0 correlation (bigger is better).*

sampled corpus containing 753×10 utterances. Each speaker contributes 735 utterances to this corpus, where 735 is the number of utterances from speaker XS01, who has the least amount of training data.

- **MU**: The conventional multi-speaker models trained with all the data from every speaker, i.e., all 32,076 training utterances from the original corpus.
- **OV**: A multi-speaker model trained with an over-sampled corpus. We used all utterances and then sampled more from minority speakers so that each got the same frequency in training. The amount of training data is 8,750×10 utterances.
- **E1**, **E2**, **E3**: Multi-speaker models trained with resampled corpora. In total, 3,000 utterances are sampled with replication from each speaker. The number of training utterances is 3,000×10, and the number of unique utterances obtained in each sampling session is listed in Table 1.
- **EN**: A non-parametric ensemble model. We simply combined the generated acoustic features obtained from the E1, E2 and E3 models using the combination functions discussed in Section 2.2.

4. Evaluations

4.1. Objective evaluations

Figure 2 shows mel-cepstral distortion between the generated and natural MGC while Fig.3 shows correlation between the generated F0 sequence inferred from the quantization output and the natural sequence. These figures show objective results separately for each speaker with color codes indicating the best system as well as the system which is better than the SD baseline. Even though objective evaluations do not directly reflect the quality of synthetic speech perceived by humans, they do demonstrate the potential of the proposed methods.

The under-sampling strategy UN with data pooled from 10 speakers does not seem to have any significant improvement over the baseline SD even for minority speaker XS01, whose entire data is included in UN. This result suggests that a multi-speaker model is not always better than the single speaker model, especially when the amount of pooled data is still limited. The over-sampling strategy OV is better than SD overall, but there is noticeable degradation in the case of majority speakers in terms of the F0 correlation metric. The conventional multi-speaker model MU shows consistent improvements over the baseline SD for most speakers. We conclude that simply pooling the data of all speakers is a reasonable strategy.

The sampling strategies E1, E2, and E3 seem to be better than the baseline SD but worse than MU. The performances vary for each session due to the stochastic nature of the sampling method. Surprisingly simply combining the generated features of E1, E2, and E3 using the average functions described in Section 2.2 produced a better result than each individual subsystem. In general, the ensemble strategy EN had the best results. Note that the amount of unique utterances from majority speakers (XL10, L09, etc.) used for the ensemble model is significant lower than the SD and MU due to the random sampling artifact, as shown in Table 1.

4.2. Subjective evaluations

We conducted a subjective listening test with samples synthesized using SD, MU and EN strategies[1]. Recorded speech is not included in our test, but we use WaveNet vocoder to synthesize

[1]Samples are available at https://nii-yamagishilab.github.io/sample-tts-speaker-imbalanced/

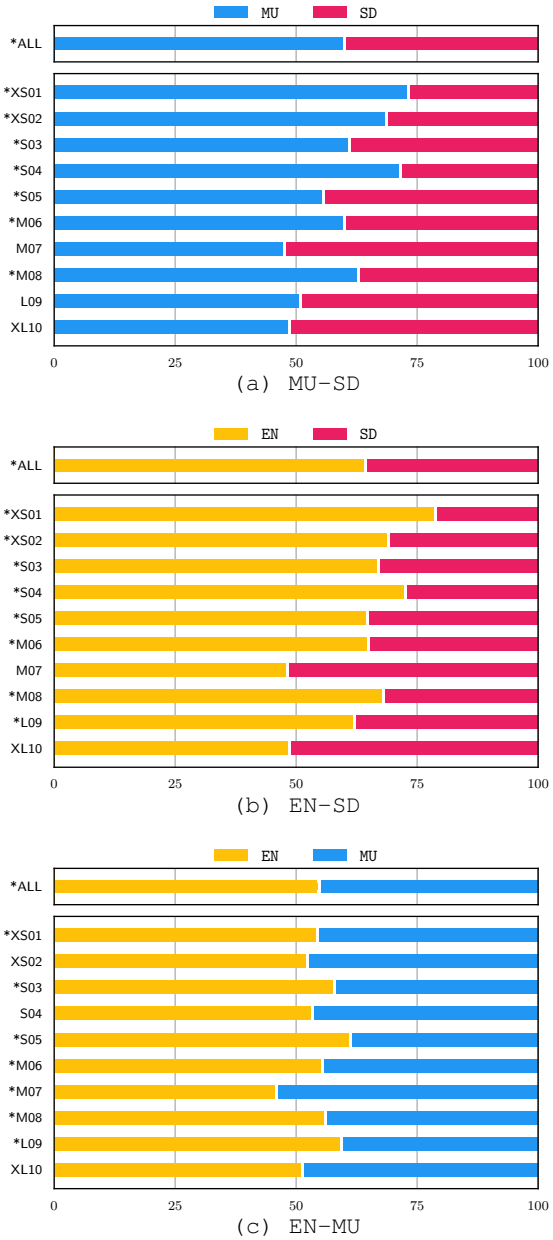

Figure 4: *AB preference test results for TTS samples of three strategies.*

speech from natural acoustic features as the reference, namely a copy synthesis strategy CO. All samples are normalized using the sv56 program. Each strategy contains 1,000 utterances, 100 utterances per speaker. We prepared a simple AB preference test in which a participant was asked to answer which sample sounds better between two presented. The presented samples are spoken by the same speaker with the same content and duration but generated from different strategies. We compared four pairs: MU−SD, EN−SD, EN−MU and the anchor test EN−CO. Each session contains one unique sentence from each of the ten target speakers, which make 40 questions in total. The question orders and sample positions are shuffled to prevent cognitive bias. Each paid participant could do ten sessions at most. We gathered answers from 997 sessions (three are discarded for incompleteness) provided by 175 participants

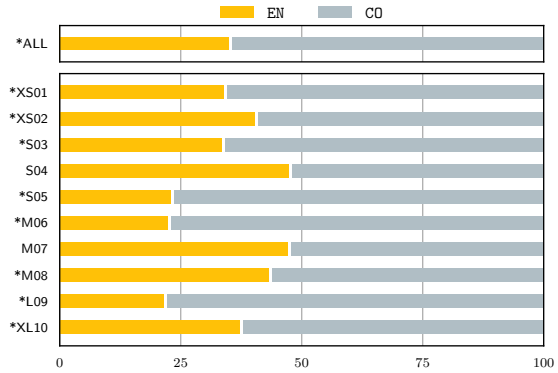

Figure 5: *Anchor AB preference test results for copy synthesis and ensemble strategy samples.*

to evaluate performance of the proposed methods. The results are calculated on both a per speaker and per strategy basis.

The preference results of the TTS samples are shown in Fig.4, where (*) indicates systems whose results are statistically significant according to the 95% confidence level of an exact binomial test. Between the multi-speaker model and single model, the result is in favor of the MU over the SD, as presented in Fig.4(a). When considering each speaker separately, we can see that speakers with less data benefit the most from the multi-speaker model, while speakers with the most data do not seem to suffer any performance degradation. A similar pattern can be seen between the ensemble model and the single model (as in Fig.4(b)), with an even stronger improvement observed with the EN strategy. Figure 4(c) shows direct comparisons between the multi-speaker model MU and the ensemble model EN. We obtained statistically significant results favoring EN for many speakers except for M07, who fared best with the MU strategy. The results of speakers XS02, S04 and XL10 while not significant, do seem to favor EN as well. To conclude, our proposed ensemble strategy showed significant improvements over the conventional multi-speaker model. The trade-off is the increased number of parameters as well as increased training and inference times due to the fact that multiple models are required. The anchor test between our proposed strategy EN and the copy synthesis CO is shown in Fig.5. As expected CO dominated, with statistically significant results for all cases except speakers S04 and M07.

5. Conclusions

We investigated the effect of a speaker-imbalanced corpus on the performance of a neural multi-speaker acoustic model. The results showed that simply combining all the available data without any resampling led to a well-rounded performance for all speakers involved. Moreover the multi-speaker model greatly benefited from a simple ensemble setup with just three subsystems sharing the same network structure but trained on different subsets of a corpus obtained through the sampling method. The one disadvantage is that the ensemble setup increases the number of parameters and the inference times. For future work, we plan to distill knowledge from an ensemble teacher network to a singular-structure student to inherit the good performance while avoiding increased parameters and processing times [23]. We also intend to introduce diversity to the network structure along with diversity in training data in order to capitalize on the strengths and reduce the weaknesses of different network structures [22].

6. References

[1] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, "Investigating accuracy of pitch-accent annotations in neural network-based speech synthesis and denoising effects," in *Proc. INTERSPEECH*, 2018, pp. 37–41.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[3] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, and L. Shen, "Corpus building for data-driven TTS systems," in *Proc. IEEE Workshop on Speech Synthesis*, 2002, pp. 199–202.

[4] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in *Proc. EUROSPEECH*, 2003.

[5] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of TTS voices trained on ASR data." in *Proc. INTERSPEECH*, 2017, pp. 3971–3975.

[6] K.-Z. Lee, E. Cooper, and J. Hirschberg, "A comparison of speaker-based and utterance-based data selection for text-to-speech synthesis," in *Proc. INTERSPEECH*, 2018, pp. 2873–2877.

[7] M. Podsiadło and V. Ungureanu, "Experiments with training corpora for statistical text-to-speech systems." in *Proc. INTERSPEECH*, 2018, pp. 2002–2006.

[8] F.-Y. Kuo, S. Aryal, G. Degottex, S. Kang, P. Lanchantin, and I. Ouyang, "Data selection for improving naturalness of TTS voices trained on small found corpuses," in *Proc. SLT*, 2018, pp. 319–324.

[9] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for hmm-based speech synthesis–analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 5, pp. 984–1004, 2010.

[10] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[11] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.

[12] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[13] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers BLSTM-RNN-based speech synthesis," in *Proc. INTERSPEECH*, 2016, pp. 2268–2272.

[14] J. Latorre, J. Lachowicz, J. Lorenzo-Trueba, T. Merritt, T. Drugman, S. Ronanki, and V. Klimkov, "Effect of data reduction on sequence-to-sequence neural TTS," in *Proc. ICASSP*, 2019, pp. 7075–7079.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res*, vol. 16, pp. 321–357, 2002.

[16] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Proc. INTERSPEECH*, 2014, pp. 1915–1919.

[17] N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE T. Inf. Syst.*, vol. 101, no. 2, pp. 462–472, 2018.

[18] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proc. ICML*, 1997, pp. 179–186.

[19] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. ICAI*, 2000, pp. 111–117.

[20] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[21] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[22] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 41–53, 2016.

[23] Y. Chebotar and A. Waters, "Distilling knowledge from ensembles of neural networks for speech recognition." in *Proc. INTERSPEECH*, 2016, pp. 3439–3443.

[24] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4895–4899.

[25] ——, "Autoregressive neural f0 model for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1406–1419, 2018.

[26] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. ASRU*, 2017, pp. 712–718.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.