*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# AISHELL-3: A Multi-Speaker Mandarin TTS Corpus

*Yao Shi*[1,2], *Hui Bu*[3], *Xin Xu*[3], *Shaoji Zhang*[3], *Ming Li*[1,2,*]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] Data Science Research Center, Duke Kunshan University, Kunshan, China
[3] Beijing Shell Shell Technology Co., Ltd, Beijing, China

ming.li369@dukekunshan.edu.cn

## Abstract

In this paper, we present AISHELL-3[†], a large-scale multi-speaker Mandarin speech corpus which could be used to train multi-speaker Text-To-Speech (TTS) systems. The corpus contains roughly 85 hours of emotion-neutral recordings spanning across 218 native Chinese mandarin speakers. Their auxiliary attributes such as gender, age group and native accents are explicitly marked and provided in the corpus. Moreover, transcripts in Chinese character-level and pinyin-level are provided along with the recordings. We also present some data processing strategies and techniques which match with the characteristics of the presented corpus and conduct experiments on multiple speech-synthesis systems to assess the quality of the generated speech samples, showing promising results. The corpus is available online at `openslr.org/93/` under Apache v2.0 license.

**Index Terms**: open source database, text-to-speech, multi-speaker speech synthesis

## 1. Introduction

Speech synthesis, or Text-to-Speech (TTS), is the automated process of mapping input text specifications to target utterances [1]. In recent years, TTS synthesis systems have achieved marvelous results in terms of audio quality and perceptual naturalness [2]. This flourishing research progress is made largely due to the introduction of neural-network based deep learning models, e.g. Tacotron [2, 3], Fastspeech [4] etc., and neural vocoders that map the lower dimensional acoustic representation to waveforms [5, 6, 7, 8].

A key characteristic of TTS is the lack of constraint, which renders the task essentially as a one-to-many mapping [1, 3]. Given only textual input, neutral or agitated speech is equally valid as output, as is speech rendered by a female or a male voice. But real-world application of such systems requires robust and consistent behaviors. This begs the question of whether we could provide further specification to the system to gain more flexibility over conventional approaches. There is a growing interest within the field in designing TTS systems that are more flexible and with stronger constraints on its behaviors. Recent publications on expressive TTS tend to associate the acoustic model with explicit control signals as augmented input besides normalized texts [9, 10, 11, 12]. A more natural and important attribute of speech is the speaker identity. And multi-speaker acoustic models give TTS systems the ability to disentangle perceptual speaker identity from the textual contents of the synthesized utterance by explicitly conditioning the model on the desired speaker [13, 14, 15, 16].

Training such systems naturally requires significant amount of annotated data. VCTK [17] is a freely available multi-speaker corpus that could be used to train such systems. However, VCTK only includes utterances in English. As suggested by previous studies [18, 19], despite the cultural influence of the English language as a *lingua franca* in the academia, language specific subsystems and model modifications are indeed an area of active research. TTS systems targeted on tonal languages such as Chinese Mandarin become more challenging given their complex tonal and prosodic structures [20]. The lack of publicly available multi-speaker TTS corpus in Mandarin makes research in this area more difficult and costly.

To this end, we introduce the freely available AISHELL-3 corpus to fill this vacancy in open resources. Furthermore, we trained multi-speaker Tacotron-2 and Fastspeech models and report the data-preparation techniques and experimental results in this paper. DiDi-Speech [21] is a concurrent work to AISHELL-3, also addressing this resource gap and contains a much larger population. However, it has less number of utterances per speaker, with relatively high level of background-noises. And it includes only character-level labels with auto-generated pronunciation annotations.

The rest of this paper is structured as follows: Section 2 introduces the presented corpus and some basic statistics. Section 3 covers the data-preparation procedure and baseline systems. Section 4 shows and interprets the experimental results obtained on the trained baseline models. Conclusion is provided in Section 5.

## 2. The AISHELL-3 Corpus

AISHELL-3 is a multi-speaker Mandarin Chinese audio corpus, which could be used to train multi-speaker TTS systems. It contains in total 88035 utterances from 218 native speakers reading texts from given scripts with neutral emotion.

Table 1 and Table 2 show the distribution of some basic attributes across the entire corpus.

### 2.1. Script Preparation and Recording

The topics of the textual scripts spread a wide range of domains including smart home voice commands, news reports, geographic information and number strings. These texts are first gathered from respective corpora. Then data-masking is performed to eliminate sensitive contents. The resulting texts are segmented by punctuation marks and further filtered to contain only Chinese-characters.

The speaker population is composed of amateur subjects covering both genders and northern/southern accents. The recording is set up in quiet indoor environments with no significant background noise or reverberation. The audio data is recorded with high-fidelity microphones (44.1kHz, 16-bit

*corresponding author

[†]samples available online at `sos1sos2sixteen.github.io/aishell3v2`



http://dx.doi.org/10.21437/Interspeech.2021-755

Table 1: *Utterance length & number per speaker distribution*

| | mean | median | max | min |
|---|---|---|---|---|
| uttr. length (character) | 11.3 | 11 | 39 | 1 |
| uttr. per speaker | 403.84 | 452 | 505 | 138 |

Table 2: *Speaker attributes distribution*

| attribute | distribution |
|---|---|
| gender | 176 female / 42 male |
| accent | 165 north / 51 south / 2 others |
| age group | $<25$: 175 / $>25$: 43 |

depth), which are 20 cm away from the speaker. All speakers read the contents of the scripts in a neutral fashion.

### 2.2. Transcription

The character-level and pronunciation-level labels included in AISHELL-3 are manually transcribed from speech after recording, reflecting the actual readings. This addresses four key difficulties in automatically deriving pinyins from Chinese character level scripts via dictionary lookups:

1.**Homograph.** Some characters could be pronounced in multiple ways depending on the textual context they resides [22]. 2.**Tone sandhi.** Some tones shift under certain phonological contexts. A good example being, normally characters in the initial part of consecutive third tones shift to the second tone, e.g. *guan3 li3* (to manage) should be pronounced as *guan2 li3*, but this rule does not apply to all such circumstances. 3.**Erization (erhua).** The Chinese character for *son* (pronounced *er2*) behaves like a normal character in some contexts, but it also acts as an rhotacization marker indicating the preceding character has an *rhotacized* final. 4.**Accents and Mispronunciations.**

These difficulties make the phonetizing process non-trivial for Chinese and the manual transcripts valuable.

## 3. System and Dataset Preparation

### 3.1. Multi-Speaker TTS Systems

To assess the feasibility and quality of the presented dataset in multi-speaker TTS tasks, we select two mainstream TTS systems, one RNN-based and one feed-forward, as the baselines. Since these two models were first published without multi-speaker support, we follow common one-hot embedding based method to perform multi-speaker TTS experiments.

#### *3.1.1. Tacotron-2*

In general, Tacotron-2 [2] is an RNN-based seq2seq structure with three major components: a CNN-BiLSTM based encoder, a LSTM decoder and an attention mechanism in between.

The location sensitive attention [23] was originally used in Tacotron-2, which leverages both contextual and locational information to determine the attention scores. However we find this formulation not generalize well on long sentences [24] and converges slower during training especially without studio-quality data. Since most of the utterances in AISHELL-3 are short, we follow [25] and adopt the GMMv2 attention, which has an open-source implementation available[1].

[1] `github.com/mozilla/TTS`

For multi-speaker synthesis, we add a 128 dimensional learnable embedding dictionary in conjunction with speaker-id labels. The embeddings are concatenated to the encoder's output sequence, conditioning the attention and decoder modules on both speaker and textual information.

#### *3.1.2. Fastspeech*

Fastspeech is a fully-feedforward architecture based on Transformer encoder-like building blocks [26]. It includes encoder and decoder modules operating on text and frame level features respectively, with a supervised duration predictor and regulator to control the correspondence between the two sequences.

Again, in order to achieve multi-speaker modeling with Fastspeech, we use trainable embedding-dictionary as speaker representation. Each embedding vector in the dictionary is 256 dimensional, matching the model dimension ($d_k$). The embedding vectors are added to the encoder output before duration prediction and decoding separately.

The feed-forward nature of Fastspeech combined with this separation simplifies the interaction between its sub-modules. This allows us to disentangle the voice and intonation of the synthesized sample, by simply injecting different embeddings in the decoder and the duration-predictor respectively. A more fluent and confident speech could be synthesized using another speakers duration embedding while still preserving the target speaker's voice. We call this inference trick cross-speaker **duration migration** (DM), and explores the impact of its use in section 4.3.

### 3.2. Dataset Preparation

We select and prepare a subset of the presented corpus as the training dataset used throughout our experiments.

#### *3.2.1. Data Partitioning*

The number of utterances per speaker in the presented corpus, though averaged to 403.84, is unevenly distributed. In order to guarantee that a more balanced training dataset is used, we selected 161 speakers as train-set speakers while the rest are reserved as testing data for textual input in evaluation.

It's worth noting that not all utterances from the train-set speakers are used in model training. Speeches containing silence segments beyond 0.4s (35 frames) are detected and kept away from training. This data filtration procedure significantly boosts the stability of the trained model.

The resulting train-set contains 56467 utterances, which is around 55 hours long.

#### *3.2.2. Duration Extraction for Fastspeech*

The duration predictor in the Fastspeech model is trained with the guidance of the alignment information produced by an auto-regressive TTS system. We use pre-trained Tacotron model to extract the duration of the training samples following [4]. However, we find the multiple heads of the GMM attention interfere with the extraction procedure. This results in highly inaccurate duration labels which causes Fastspeech to produce blurred outputs.

We therefore constrained the $\arg\max$ operator in the originally proposed extractor to operate on a small window sized $2w$ centered around last time-step's spotted alignment region $c_{i-1}$. For attention map $a \in \mathbb{R}^{S\times T}$, and desired duration sequence $\{d_i\} \in \mathbb{N}^T$, where $S$ denotes the length of the spectrogram and $T$ the text sequence. We extract the duration sequence

Table 3: *Dataset usage*

| Model | Dataset |
|---|---|
| Tacotron-2 | AISHELL-3 |
| Fastspeech | AISHELL-3 |
| HiFi-GAN | fine-tuned on AISHELL-3 |
| ecapa-tdnn | vox2 [27], tuned on AISHELL-2 [28] |
| resnet-se | private dataset including AISHELL-3 |

with Eq. (1) and (2). This improves the extraction accuracy and enables us to train Fastspeech with minimum front-end preprocess.

$$d_i = \Sigma_{s=1}^{S}[c_s = i] \tag{1}$$

$$\begin{cases} c_1 = 0, \\ c_i = \arg\max_{|c_{i-1}-t|<w} a_{i,t}, \end{cases} \tag{2}$$

# 4. Experimental Results

We implement and train the baseline TTS systems using the proposed corpus and perform evaluations on the synthesized samples. The experimental setup and results are described in this section.

## 4.1. Experimental Setup

The Tacotron-2 and Fastspeech models are trained using Adam optimizer with $(lr = 10^{-3}, weight\_decay = 10^{-6})$. We use a batch-size of 32 in our experiments. Tacotron and Fastspeech converge at $50k$ and $400k$ steps respectively. But given Fastspeech's fast iteration speed, the two takes approximately the same amount of time to train.

We generate all our test-samples using HiFi-GAN [8]. The model is based on an open-source pre-trained checkpoint[2], and fine-tuned with GTA features generated by Tacotron. Table 3 summarizes the dataset usage in our experiments.

## 4.2. MOS on Naturalness

We synthesize 10 samples for each trained system to perform MOS naturalness evaluations. The evaluation is conducted by 20 native mandarin speakers, and all samples (including ground-truths, which are down-sampled) are in 22kHz format. The results are shown in Table 4, where **real** speakers denote target speakers from the trained embedding table, and **sampled** speakers are unseen new embedding vectors sampled from a distribution (see section 4.5 for experimental procedure and analysis).

The overall results show that both trained systems are able to produce high-quality speeches from textual inputs. However, the entry for vocoder analysis-synthesis samples perform worse than nearly all synthesis systems, which is irregular. This may be due to the vocoder used is fine-tuned directly with Tacotron-GTA features, causing it to produce obvious artifact when used with ground-truth spectrograms.

## 4.3. Impact of Cross-Speaker Duration Migration

An unfortunate consequence of using speech corpus recorded by amateur subjects is that some speakers are unsure of their scripts and present a hesitant or mechanical tone. these attributes can be captured by the trained acoustic model and may degrade the overall naturalness as perceived by listeners.

[2] `github.com/jik876/hifi-gan` UNIVERSAL_V1

Table 4: *MOS results*

| System | Speakers | MOS (95% CI) |
|---|---|---|
| gt | - | $4.51(\pm0.09)$ |
| vocoded | - | $4.01(\pm0.13)$ |
| Tacotron | real | $4.21(\pm0.09)$ |
| Fastspeech | | $4.08(\pm0.11)$ |
| Tacotron | sampled | $3.89(\pm0.11)$ |
| Fastspeech | | $4.14(\pm0.11)$ |

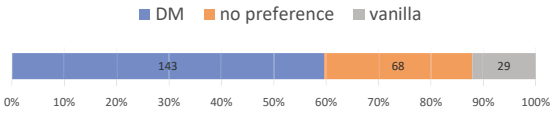


Figure 1: *Prosody preference test result*

The structure of the Fastspeech model provides a way of replacing "bad" duration predictions with fluent ones during inference (DM). We conduct a subjective preference test to explore the impact of DM on listeners' perceptual preference in identical models. 4 speakers exhibiting said flaws are picked as target voices, and a single confident speaker as target duration. We synthesized 12 pairs of utterances with the same content and voice but different durations. 20 listeners are invited to determine which one is more preferable prosody-wise. The result is illustrated in Figure 1.

This result suggests that cross-speaker duration migration is preferred in such cases where the speaker's intonation is less than ideal. The impact of this strategy in terms of speaker similarity is also studied in section 4.4, which shows that the target's voice is well preserved under DM.

## 4.4. Objective Evaluation of Speaker Similarity

We conduct objective evaluation regarding speaker similarities on synthesized speech samples. The intention of this experiment is to assess the ability of the TTS models to synthesize speeches with the target speakers' voice. We employ two separately trained Speaker Verification (SV) systems as judges of speaker similarities. The SV systems are marked as *resnet-se* [29] and *ecapa-tdnn* [30]. Also note that AISHELL-3 is included in the training data of *resnet-se*. Thus it is an in-domain evaluation for *resnet-se*.

Given an SV system, the speaker similarity between two speech samples can be measured by the **cosine similarity** of the embedding vectors extracted from the SV model. Therefore, the *compactness* of a group of embeddings could be characterized by the intra-class similarity ($intra$), and conversely, *well-separatedness* by inter-class similarity ($inter$).

The **Equal Error Rate** (EER) is a widely adopted metric in literature as a performance indicator of SV systems. In addition, it can be used as a measurement for the quality of a multi-speaker speech synthesis system, which aims to produce results that are indistinguishable from real recordings in terms of speaker similarity [14, 16].

For every speaker from the training set, 15 utterances are drawn from the corpus as ground-truths (gt). Correspondingly 15 samples are generated using the trained multi-speaker TTS systems. Trials for EER measurement are generated by sampling $10^4$ (enroll, verify) pairs from a pool $E_{gt} \times (E_{gt} \cup E_{syn})$, where $E_{gt}$ stands for the set of ground-truth embeddings, and $E_{syn}$ the set of synthesized embeddings.

Table 5: *Speaker similarity evaluation results*

| System SV | System TTS | Cosine Similarity *intra* | Cosine Similarity *inter* | EER(%) |
|---|---|---|---|---|
| ecapa-tdnn | gt | 0.900 | 0.371 | 1.07 |
| | tacotron | 0.827 | 0.306 | 1.43 |
| | fastspeech | 0.839 | 0.311 | 1.46 |
| | fast+DM | 0.829 | 0.301 | 1.51 |
| resnet-se | gt | 0.884 | 0.037 | 0.06 |
| | tacotron | 0.746 | 0.035 | 0.34 |
| | fastspeech | 0.739 | 0.038 | 0.36 |
| | fast+DM | 0.731 | 0.036 | 0.37 |

Results concerning speaker similarity are shown in Table 5, where the mean intra-/inter-class similarity over all speakers are reported along with the EER. The similarity measurements from both SV systems show a drop between ground-truths and synthesis systems, and a mild downward trend across the three TTS models as the model architecture becomes more constrained. However, the highest EER $1.51\%$ from our evaluation results still can be considered as an outstanding performance on speaker similarity. This implies that voices from seen speakers can be stably reproduced by the trained models.

### 4.5. Generalization Potential

The embedding dictionary based approach we employ limits the known target voices to be exactly the speakers in the train-set, as the corresponding relationship is built along-side model training. But synthesizing new voices is still possible under this setup. One way is to sample embeddings from the Gaussian space estimated from trained embeddings. Nevertheless, the quality and variety of the voices obtained through this method are not guaranteed. For example, $N$ distinct sampled embeddings may correspond to only one voice. We conduct experiments to explore the generalization potential and the quality of samples synthesized using this method.

We fit the trained embedding table of each TTS system using a 3-component full-variance Gaussian Mixture Model following [31] and randomly sample $N = 4000$ vectors as the set of potential speakers $S$. We synthesize $n = 10$ utterances for each speaker to form the sample set $U$. Then 10 utterances are randomly drawn from $U$ for MOS evaluation to assess the quality of sampled voices. We then try to determine the variety of potential voices by counting the number of distinct speakers in $U$.

Again, we extract speaker embeddings $e_{s,i}$ for each utterance in $U$, and calculate the mean embedding per speaker $\{\bar{e}_s\}$ as well as the cross-speaker similarity matrix $\mathbf{D}$.

$$\bar{e}_s = \frac{1}{n}\Sigma_{i=1}^{n} e_{s,i} \tag{3}$$

$$\mathbf{D}_{i,j} = cos(\bar{e}_i, \bar{e}_j) \tag{4}$$

$$\mathbf{A} = \mathbf{D} < d \tag{5}$$

Since the speaker population is over-sampled from a system of 161 known speakers, we expect a smooth continuous transition among sampled speakers in $\{\bar{e}_s\}$. Therefore, an unsupervised clustering such as DBSCAN [32] does not yield the correct approximation. Instead, we define two speakers $i, j$ are *mutually exclusive* if $D_{i,j} < d$ for a preset threshold $d$. The threshold $d$ is determined as the $5\%$ quantile in the distribution of the distance between known sample points and their respective class-mean. Under this formulation, the boolean matrix $\mathbf{A}$ in Eq. (5) can be considered as the adjacency matrix for an undirected graph $G$ describing the mutually exclusive relationship. And the number of distinct speakers equals the number of vertices ($|V_c|$) in the maximal complete subgraph (clique) of $G$. Since finding maximal clique is NP-complete, we employ multiple greedy search from random sources and report the maximum in Table 6.

Table 6: *Speaker generalization potential*

| System SV | System TTS | Generalization $d$ (5%) | Generalization $\lvert V_c\rvert$ |
|---|---|---|---|
| ecapa-tdnn | tacotron | 0.823 | 3381 |
| | fastspeech | | 648 |
| resnet-se | tacotron | 0.791 | 3724 |
| | fastspeech | | 1579 |

This number characterizes the span of potential distinct voices the trained model could generalize to. This reflects the variety of the underlying dataset seen during training. But the results may not be interpreted as 1.the maximum number of potential speakers, since the clique found is not the global maximal, and it is limited by $N$; 2. an enumeration of actual speakers unknown to us, since the clique found is not unique.

The experimental results show great generalization potential given all $|V_c|$ greatly exceeds the number of speakers from our train-set (161). We note also that Fastspeech have a lower $|V_c|$ count than Tacotron, but give a higher MOS score in Table 4. We found that 3 samples included in scoring for Tacotron includes unusual raising pitch at the end. By manually eliminating the scores for these samples, MOS for entry tacotron+sampled becomes $4.10(\pm 0.11)$. Though this exclusion is irregular, it hints the potential of the presented corpus on more stable systems.

## 5. Conclusions

In this paper, a new publicly available Mandarin speech corpus that could be used for multi-speaker TTS systems is presented. Two representative TTS systems are trained and evaluated to highlight the overall quality of the corpus. We evaluate the performance of the resulting models subjectively in terms of overall MOS scores and listener preferences. Furthermore, objective evaluations are conducted with regards to speaker similarity and generalization capacity which reflects the variety of the underlying dataset. We found that our trained systems are capable of producing speeches with decent quality (with MOS up to $4.21$), and even improved prosody in certain cases. Moreover, experiments show the trained models generalize well to a potentially vast range of voices. We believe that the presented corpus is valuable for TTS research in Mandarin, and we see vast opportunities for more sophisticated approaches to improve upon the reported baselines.

## 6. Acknowledgements

This research is funded in part by the National Natural Science Foundation of China (61773413), the Fundamental Research Funds for the Central Universities (2042021kf0039), Key Research and Development Program of Jiangsu Province (BE2019054), Science and Technology Program of Guangzhou City (201903010040,202007030011) and Six Talent Peaks Project in Jiangsu Province (JY-074)

# 7. References


[1] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. of ICASSP 2018*, pp. 4779–4783.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. of Interspeech 2017*, pp. 4006–4010.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in NeurIPS 2019*, pp. 3165–3174.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop*, 2016, p. 125.

[6] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. of ICASSP 2019*, pp. 3617–3621.

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in NeurIPS 2019*, pp. 14 910–14 921.

[8] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in NeurIPS 2020*.

[9] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. of ICASSP 2020*, pp. 6189–6193.

[10] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. of ICML 2018*, vol. 80, pp. 5167–5176.

[11] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. of ICASSP 2020*, pp. 6264–6268.

[12] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *Proc. of ICASSP 2020*, pp. 6699–6703.

[13] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in NIPS 2017*, pp. 2962–2970.

[14] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in NeurIPS 2018*, pp. 4480–4490.

[15] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. of ICASSP 2020*, pp. 6184–6188.

[16] Z. Cai, C. Zhang, and M. Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," in *Proc. of Interspeech 2020*, pp. 3974–3978.

[17] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh, The Centre for Speech Technology Research*, 2016.

[18] H. Pan, X. Li, and Z. Huang, "A mandarin prosodic boundary prediction model based on multi-task learning." in *Proc. of Interspeech 2019*, pp. 4485–4488.

[19] Y. Yan, J. Jiang, and H. Yang, "Mandarin prosody boundary prediction based on sequence-to-sequence model," in *Proc. of ITNEC 2020*, vol. 1, pp. 1013–1017.

[20] N. Minematsu, S. Kobayashi, S. Shimizu, and K. Hirose, "Improved prediction of japanese word accent sandhi using CRF," in *Proc. of Interspeech 2012*, pp. 2562–2565.

[21] T. Guo, C. Wen, D. Jiang, N. Luo, R. Zhang, S. Zhao, W. Li, C. Gong, W. Zou, K. Han *et al.*, "Didispeech: A large scale mandarin speech corpus," *CoRR*, vol. abs/2010.09275, 2020. [Online]. Available: https://arxiv.org/abs/2010.09275

[22] Z. Cai, Y. Yang, C. Zhang, X. Qin, and M. Li, "Polyphone Disambiguation for Mandarin Chinese Using Conditional Neural Network with Multi-Level Embedding Features," in *Proc. Interspeech 2019*, pp. 2110–2114.

[23] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in NIPS 2015*, pp. 577–585.

[24] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of an encoder-decoder end-to-end TTS framework using marginalization of monotonic hard latent alignments," *CoRR*, vol. abs/1908.11535, 2019. [Online]. Available: http://arxiv.org/abs/1908.11535

[25] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. of ICASSP 2020*, pp. 6194–6198.

[26] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. of AAAI 2019*, vol. 33, pp. 6706–6713.

[27] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. of Interspeech 2018*, pp. 1086–1090.

[28] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: transforming mandarin ASR research into industrial scale," *CoRR*, vol. abs/1808.10583, 2018. [Online]. Available: http://arxiv.org/abs/1808.10583

[29] J. S. Chung, J. Huh, S. Mun, M. Lee, H. Heo, S. Choe, C. Ham, S. Jung, B. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. of Interspeech 2020*, pp. 2977–2981.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech 2020*, pp. 3830–3834.

[31] Y. Huang, Y. Chen, J. Pelecanos, and Q. Wang, "Synth2aug: Cross-domain speaker recognition with tts synthesized speech," in *2021 IEEE SLT*, pp. 316–322.

[32] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.