applied sciences

MDPI

Article

Using Transfer Learning to Realize Low Resource Dungan Language Speech Synthesis

Mengrui Liu [1,†], Rui Jiang [2,†] and Hongwu Yang [2,3,*]

[1] College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China; liuxh709@163.com
[2] School of Educational Technology, Northwest Normal University, Lanzhou 730070, China; jiangh940618@163.com
[3] Key Laboratory of Education Digitalization of Gansu Province, Lanzhou 730070, China
* Correspondence: yanghw@nwnu.edu.cn
† These authors contributed equally to this work.

Abstract: This article presents a transfer-learning-based method to improve the synthesized speech quality of the low-resource Dungan language. This improvement is accomplished by fine-tuning a pre-trained Mandarin acoustic model to a Dungan language acoustic model using a limited Dungan corpus within the Tacotron2+WaveRNN framework. Our method begins with developing a transformer-based Dungan text analyzer capable of generating unit sequences with embedded prosodic information from Dungan sentences. These unit sequences, along with the speech features, provide <unit sequence with prosodic labels, Mel spectrograms> pairs as the input of Tacotron2 to train the acoustic model. Concurrently, we pre-trained a Tacotron2-based Mandarin acoustic model using a large-scale Mandarin corpus. The model is then fine-tuned with a small-scale Dungan speech corpus to derive a Dungan acoustic model that autonomously learns the alignment and mapping of the units to the spectrograms. The resulting spectrograms are converted into waveforms via the WaveRNN vocoder, facilitating the synthesis of high-quality Mandarin or Dungan speech. Both subjective and objective experiments suggest that the proposed transfer learning-based Dungan speech synthesis achieves superior scores compared to models trained only with the Dungan corpus and other methods. Consequently, our method offers a strategy to achieve speech synthesis for low-resource languages by adding prosodic information and leveraging a similar, high-resource language corpus through transfer learning.

Keywords: Dungan language speech synthesis; text analysis; transfer learning; low-resource language; tacotron2

Citation: Liu, M.; Jiang, R.; Yang, H. Using Transfer Learning to Realize Low Resource Dungan Language Speech Synthesis. *Appl. Sci.* **2024**, *14*, 6336. https://doi.org/10.3390/app14146336

Academic Editors: Gloria Corpas Pastor and Tharindu Ranasinghe

Received: 17 June 2024
Revised: 17 July 2024
Accepted: 18 July 2024
Published: 20 July 2024

1. Introduction

Speech synthesis (text-to-speech (TTS) conversion) is widely used in smart homes, navigation systems, and audiobook applications. Globally, there are approximately 6000 languages, most considered low-resource. While significant progress has been made in speech synthesis for major languages like Mandarin, English, and French, the voice quality of TTS for low-resource languages, such as Tibetan and Dungan, remains suboptimal. In recent years, there has been a surge in research focused on low-resource language speech synthesis, as evidenced by numerous studies [1–6]. However, research on Dungan language speech synthesis still needs to be completed. The Dungan language, which is a variant of the Shanxi-Gansu dialects within the Chinese dialect spoken in Central Asia, is classified as a low-resource language due to its limited usage, dwindling number of speakers, and scarcity of linguistic materials [7,8]. Given that Russian has become the official language of Central Asia, creating a comprehensive speech corpus with linguistic knowledge for high-quality Dungan speech synthesis presents a significant challenge. Although we implemented a

DNN-based Dungan speech synthesis [9,10], the synthesized speech quality was not high due to the limited training corpus.

Speech synthesis technologies encompass unit selection-based concatenative speech synthesis [11], hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) [12], and deep learning-based speech synthesis [13,14]. While deep learning has significantly advanced speech synthesis technology, methods such as long short-term memory (LSTM) and bidirectional LSTM [15,16] have addressed temporal information limitations. Moreover, end-to-end speech synthesis models [17] like Tacotron [18] and Tacotron2 [19] have demonstrated the ability to map text directly to speech. When trained with large-scale text-to-speech pairs, these models produce synthesized speech using high-quality vocoders such as the Griffin-Lim algorithm [20], WaveNet [21], and WaveRNN [22]. However, such systems require substantial training corpora. For low-resource languages, the lack of training corpus makes it difficult for end-to-end models to learn the prosodic structure of sentences, resulting in a lack of prosodic changes in synthesized speech, which affects its naturalness, posing challenges for speech synthesis of low-resource languages.

Cross-language transfer learning [23–25] has been employed to mitigate the issue of insufficient training corpora for speech synthesis in low-resource languages. This technique entails training a language model using a combination of a large corpus from a high-resource language and a smaller corpus from a low-resource language, followed by adapting this model to the low-resource language. Transfer learning in speech synthesis has proven to be an effective strategy for producing speech in low-resource languages by harnessing the capabilities of a high-resource language acoustic model [26,27].

In our prior research on Tibetan speech synthesis [28–32], we determined that integrating prosodic information through transfer learning-based techniques enhances the quality of synthesized speech for low-resource languages such as Tibetan. Building on this insight, the present study implements a sequence-to-sequence (seq2seq) approach for Dungan language speech synthesis, leveraging transfer learning and prosodic information within the Tacotron2+WaveRNN framework. This method entails utilizing a Dungan text analyzer to extract prosodic labels from Dungan sentences for model integration, employing a Tacotron2-based Mandarin acoustic model, and fine-tuning the Dungan language acoustic model with a limited Dungan speech corpus. The primary contributions are delineated below:

- Front-end: We have implemented a complete text analyzer for the Dungan language, encompassing modules for text normalization, word segmentation, prosodic boundary prediction, and unit generation based on transformer technology. This analyzer can produce initials and finals as speech synthesis units with prosodic labels from Dungan sentences.
- Back-end: We have achieved seq2seq Dungan language speech synthesis by adapting a pre-trained Mandarin acoustic model within the Tacotron2+WaveRNN framework. This was accomplished by replacing Tacotron2's location-sensitive attention with forward attention, enhancing convergence speed and stability.

The rest of the article is organized as follows. We first present our transfer learning-based Dungan speech synthesis framework under Tacotron2+WaveRNN in Section 2. The experimental setup and results are presented in Section 3, while the results are discussed in Section 4. Finally, a brief conclusion and outline for future work are provided in Section 5.

2. Models and Methods

The proposed framework of transfer learning-based low-resource Dungan speech synthesis, which is shown in Figure 1, including a feature extraction module, a pre-trained Mandarin acoustic model, a transfer learning-based Dungan acoustic model training module, and a WaveRNN vocoder-based speech synthesizer.

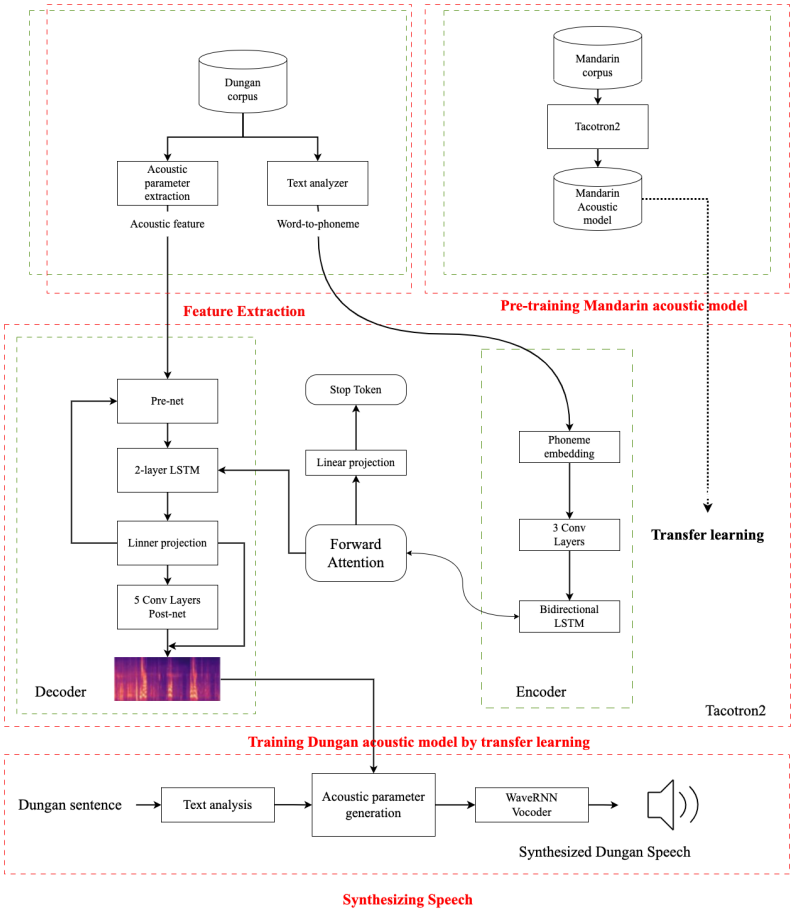

Figure 1. The framework of Tacotron2+WaveRNN-based Dungan speech synthesis.

The feature extraction module extracts acoustic features such as Mel Spectrogram from speech signals and speech synthesis units sequence from sentences. We have developed a complete Dungan language text analyzer to extract speech synthesis units with prosodic features to map Dungan sentences onto unit sequences. Given that both the Mandarin and Dungan languages utilize initials and finals as their core speech synthesis units, the resulting unit sequence incorporates these elements and pertinent prosodic information, including syllable tones and sentence-level prosodic boundary labels.

Since Tacotron2 is one of the most popular encoder-to-decoder speech synthesis frameworks, and the WaveRNN vocoder can generate natural speech, we use Tacotron2 to train acoustic models and WAVRNN to convert spectrogram to waveform for both the Dungan language and Mandarin. The Mandarin acoustic model is pre-trained with a large-scale Mandarin corpus, while the Dungan language model is transferred from the Mandarin acoustic model with a small-scale Dungan corpus.

In the speech synthesis stage, the WaveRNN vocoder generates Dungan or Mandarin speech from the input of Dungan or Chinese sentences. The text analyzer first generates the context-dependent labels from the input sentence. Then, the speech synthesis unit sequences (initials and finals with their prosodic information) are fed into the Mandarin or Dungan acoustic model to generate the Mel spectrogram. The WaveRNN vocoder is finally used to generate the speech waveforms from the Mel spectrogram. We use a home-grown Chinese text analyzer for Chinese text analysis.

2.1. Text Analyzer of Dungan Language

Unlike the prevalent seq2seq speech synthesis techniques designed for major languages that solely utilize the <phoneme sequence, speech> pair for training acoustic models, our approach employs a unit sequence incorporating prosodic labels such as the

tone of each syllable and the prosodic boundary of a sentence, serving as the "phoneme sequence". Consequently, it becomes essential to devise a comprehensive text analyzer capable of extracting a sentence's unit sequences and their prosodic labels. To this end, leveraging our in-house Chinese text analyzer, we developed a Dungan language text analyzer, as illustrated in Figure 2. The process begins with normalizing and segmenting the input Dungan sentence to determine the word boundary. A prosodic boundary analysis follows this to identify both the prosodic word and prosodic phrase boundary. In the final stage, the initials and finals of Dungan characters are derived through a transformer-based characters-to-unit conversion process.

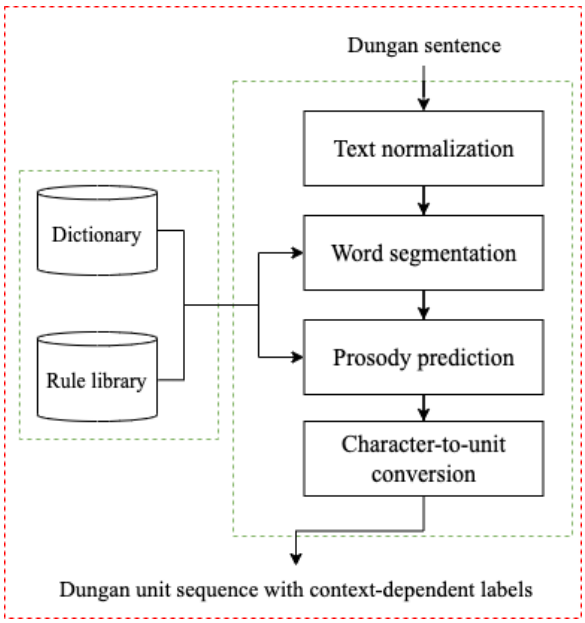

Figure 2. Procedure of Dungan text analysis.

2.1.1. Speech Synthesis Unit of Dungan language

Despite utilizing a different writing system, Dungan represents a dialectal pronunciation of Mandarin outside of China. The Dungan language is written in Cyrillic script, resembling Slavic languages like Russian, so the Dungan language is phonetic characters with sequential spelling, following a structure similar to Chinese [33–35]. The spelling order for Dungan characters consists of initials, finals, and tone, as depicted in Figure 3.

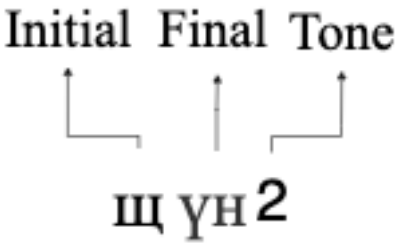

Figure 3. Structure of a Dungan character.

This article uses initials and finals as the speech synthesis unit. The Dungan character comprises 25 initials (including the zero initial) and 32 finals, as shown in Table 1. Like Mandarin, the Dungan language's tones are crucial in distinguishing semantics and emotions [36]. Dungan features four tones, excluding the light tone, namely the level tone (21), rising tone (24), falling-rising tone (53), and falling tone (44), each denoted by the numbers 1 to 4, respectively.

Table 1. The initials and finals of Dungan Language.

initials	/b/, /p/, /m/, /f/, /v/, /z/, /c/, /s/, /d/, /t/,/n/,/l/ /zh/, /ch/,/sh/, /r/, /j/, /q/, /x/, /g/, /k/, /ng/, /h/, /ϕ/
finals	/ii/, /iii/, /i/, /u/, /y/, /a/, /ia/, /ua/, /e/, /ue/, /ye/, /iE/ /ap/, /ai, /uai/, /ei/, /ui/, /ao/, /iao/, /ou/, /iou/, /an/, /ian/ /uan/, /yan/, /aN/,/iaN/, /uaN/, /uN/, /iN/, /yN/

2.1.2. Text Normalization

Any input sentence may contain numerical forms of time, date, abbreviations, and special proprietary nouns. Before converting a sentence into a sequence of phonetic symbols, it is essential to use text normalization to transform non-standard text into a unified phonetic symbol. Therefore, we implemented a rule-based text normalization to identify non-Dungan characters. We developed a set of Dungan text normalization rules based on Chinese text normalization rules [37] and employed the add-restore method to normalize the Dungan characters according to [38].

2.1.3. Word Segmentation

Word boundaries play a significant role in predicting prosodic boundaries. Thus, it is essential to identify the word boundaries of a sentence post-normalization. Dungan sentences exhibit clear distinctions between words and syllables, making segmentation relatively straightforward. We employed a maximum matching-based word segmentation algorithm to extract Dungan words from the input sentence. We compiled a Dungan word dictionary comprising 49,293 words to facilitate this process. The longest word spans eight characters in this dictionary, while the shortest is a single character. The dictionary primarily encompasses core Dungan terms, as referenced in sources like "Common Dictionary of Dungan Language" [39], "A Survey on Tungan Language in Central Asia" [40], "A Survey of Dungan Language" [41], and additional searchable Dungan terms available online.

2.1.4. Prosodic Boundary Prediction

Our approach utilizes initials and finals, along with their prosodic labels, as the input unit sequence for the acoustic model. Thus, extracting the prosodic structure from Dungan sentences is crucial for synthesizing high-quality speech. Like Mandarin, Dungan's prosodic hierarchy can be segmented into prosodic words, prosodic phrases, intonation phrases, and sentence pauses. The boundary of intonation phrases can be easily identified using Dungan punctuation marks. In this study, we employed a BiLSTM with a conditional random field (BiLSTM_CRF)-based method, as illustrated in Figure 4, to predict the boundaries of prosodic words and phrases [42].

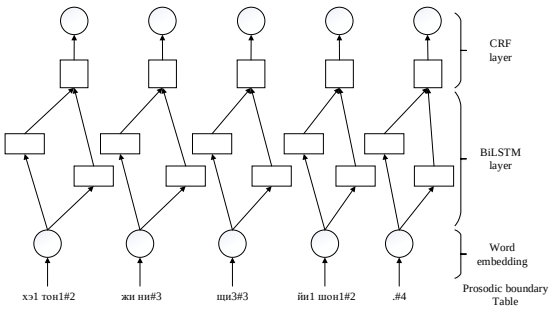

Figure 4. The framework of BLSTM_CRF-based Dungan Prosodic Boundary Prediction. The input is a Dungan sentence with prosodic information.

We employed four distinct prosodic word-position labeling sets (#1, #2, #3, #4) to categorize Dungan words into prosodic phrases. Specifically, #1 was utilized to denote the prosodic words, #2 designated the prosodic phrases, #3 marked the termination of a Dungan

word, and #4 indicated a pause within a sentence. The labeling process incorporated phrase and prosodic information derived from Dungan text, which was tagged manually. During this phase, linguists sporadically reviewed and amended selected sentences. We attained a high level of consistency with linguistic experts through iterative corrections.

Despite the BiLSTM's capability to learn context-dependent information, its independent classification decisions are constrained by strong dependencies across the output label. To address this, we employ a CRF layer that considers neighboring tags, as illustrated in Figure 4. For a normalized input sentence $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ containing n words and a tag sequence of sentence $y = (y_1, y_2, \ldots, y_n)$, each word is represented as a d-dimensional vector by word2vec. We define its prediction score $s(\mathbf{X}, y)$ to be as follows:

$$s(\mathbf{X}, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} A_{y_i,y_{i+1}} \tag{1}$$

where $\mathbf{P}$ is the matrix of scores output by the BLSTM network. P_{i,y_i} corresponds to the score of the y_i tag of the ith word in a sentence. $\mathbf{A}$ is the transition scores matrix of the CRF layer, and $A_{y_i,y_{i+1}}$ corresponds to the score from tag y_i to tag y_{i+1}.

In the training, we maximize the following log-likelihood functions:

$$\log(p(y \mid \mathbf{X})) = s(\mathbf{X}, y) - \log\left(\sum_{\widehat{y} \in \mathbf{Y_X}} e^{s(\mathbf{X}, \widehat{y})}\right) \tag{2}$$

where $\mathbf{Y_X}$ represents all possible tag sequences for an input text $\mathbf{X}$.

In the decoding, the optimal sequence y^* is given as follows:

$$y^* = \underset{\widetilde{y} \in \mathbf{Y_X}}{\operatorname{argmax}} s(\mathbf{X}, \widehat{y}) \tag{3}$$

2.1.5. Transformer-Based Character-to-Unit Conversion

Mandarin and Dungan employ the same Pinyin system for pronunciation labeling. Consequently, the character-to-unit conversion in Dungan parallels that of Mandarin. This study introduces a transformer-based approach [43] to derive the Dungan unit, as illustrated in Figure 5, to enhance the accuracy of Dungan character-to-unit conversion. The encoder and decoder are formed by stacking the same essential layers with $N = 6$. Each underlying layer consists of two sublayers. The first sublayer is the multi-head attention layer. The decoder has a layer of hidden multi-head attention (masked multi-head attention).

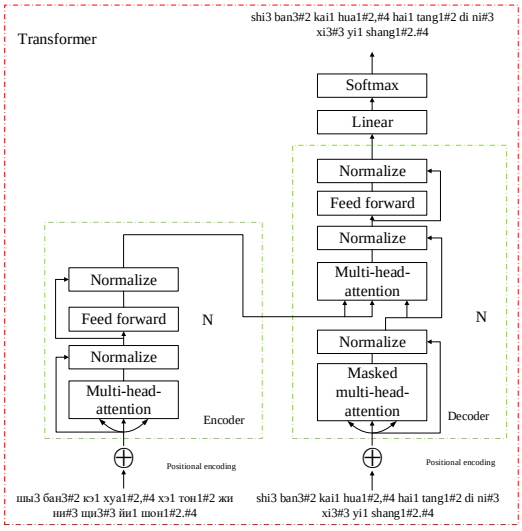

Figure 5. The framework of Transformer-based Dungan character-to-unit conversion. The input is a Dungan sentence with prosodic information (**left**) and its corresponding Pinyin sequence (**right**). The output is the Pinyin sequence with prosodic information.

2.2. Transfer Learning-Based Dungan Acoustic Model

We implement the Dungan acoustic model by fine-tuning a pre-trained Mandarin acoustic model, as illustrated in Figure 6.

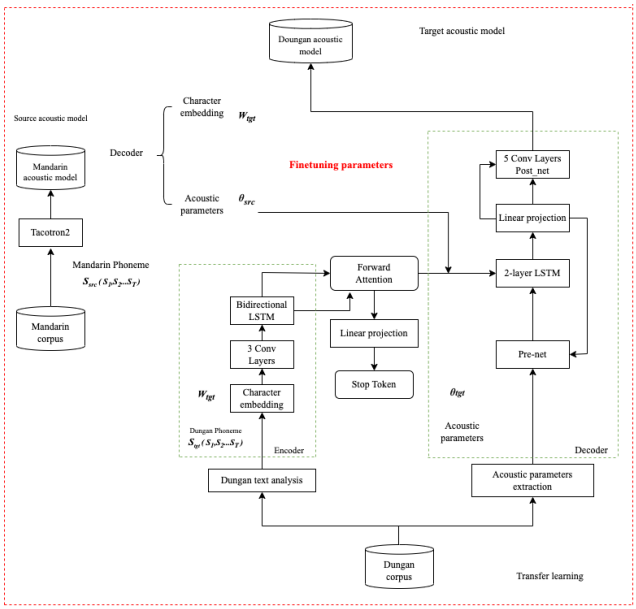

Figure 6. Procedure of training the Dungan language acoustic model with transfer learning.

2.3. Pre-Trained Tacotron2-Based Mandarin Acoustic Model

The Mandarin acoustic model is initially trained using a large-scale Mandarin corpus. Our proprietary Chinese text analyzer extracts these sentences' initials, finals, and associated prosodic labels. The extracted acoustic features encompass the Mel spectrogram from the large-scale Mandarin corpus within the Tacotron2 framework.

Given the similar pronunciation between the Dungan language and Mandarin, we employ the mapping-transfer learning method [44] to obtain a Dungan (target language) acoustic model by transferring knowledge from Mandarin (source language), which can be formulated as follows:

$$f_{\theta,W} : \mathcal{X}_{\mathcal{L}} \rightarrow \mathcal{Y} \tag{4}$$

where θ is the parameters of the acoustic model, W denotes learnable symbol embeddings, and $\mathcal{Y}$ represents the space of Mandarin. $\mathcal{X}_{\mathcal{L}}$ is the text space for the Dungan language.

$$\mathcal{X}_{\mathcal{L}} = \left\{ \{s_t\}_{t=1}^{T} \mid \forall t s_t \in \mathcal{L}, T \in \mathbb{N} \right\} \tag{5}$$

where $\mathcal{L}$ is the unit set for the Dungan language, S_t is the t-th unit of Dungan unit sequence, and T is the length of the unit sequence.

In the encoder, we input a Dungan unit sequence represented by character embeddings. This is passed through a stack of three convolutional layers, followed by batch normalization and ReLU activations. Subsequently, the output from the final convolutional layer is fed into a bidirectional LSTM layer to generate the Dungan unit features.

Mapping-based transfer learning involves mapping instances from θ_{src} and θ_{tgt} into a new acoustic parameter space. In this process, we can directly use W_{src} and θ_{src} decoded from the Mandarin acoustic model by the decoder. θ_{src} and θ_{tgt} can take embeddings as input and generate speech. However, because s_{src} and s_{tgt} come from different symbol sets, i.e., $\mathcal{L}_{src} \neq \mathcal{L}_{tgt}$, the same concept cannot be directly applied to W_{src} and W_{tgt}. To address this issue, Dunggan units are embedded in W_{tgt} to facilitate relearning during the transmission process.

We adopt the forward attention mechanism, which uses cumulative attention weights to calculate the context vector.

The decoder is an autoregressive recurrent neural network that predicts a θ_{tgt} from the encoder input Dungan unit sequence one frame at a time. We can use θ_{src} learned from the Mandarin acoustic model to initialize θ_{tgt} in the new acoustic parameter space. The output from the initial timestep is first processed through a Pre-net consisting of two fully connected layers. This output is combined with the forward attention context vector and passed through a pair of LSTM layers. The combination of the LSTM outputs and attention context vectors undergo three distinct linear transformations to predict the target spectrogram frame, stop token, and estimated residual. Subsequently, the predicted acoustic features are subjected to five convolutional layers, generating a residual to enhance the reconstruction of the Dungan acoustic model.

3. Results

3.1. Evaluation on Transformer-Base Dungan Character-to-Unit Conversion

The text analysis in the front end affects the quality of speech synthesis in the back end, so we evaluated the Dungan text analyzer, where the character-to-unit conversion module is the most critical factor affecting the quality of synthesized speech. To assess the viability of the transformer-based Dungan character-to-unit conversion module, we utilized a dataset comprising 10,783 sentences in the Dungan language transcribed using Mandarin Pinyin. The dataset's Dungan language and Mandarin Pinyin representations are isomorphic, encapsulating textual attributes like tone and prosodic boundaries inherent to the Dungan language. In our research, we allocated 10% of the total 10,783 sentences to serve as the test set, another 10% as the validation set, and the remaining 80% were designated as the training set. The hyperparameters associated with the Transformer are detailed in Table 2. We employed precision, recall, and F1 measures as evaluation indices, as illustrated in Table 3. The outcomes from the evaluation process affirmed that the proposed Dungan character-to-unit module is suitable for subsequent speech synthesis evaluation.

Table 2. The hyperparameters of Transformer-based Character-to-unit conversion model.

Parameter	Value
Attention layers N_x	6
Heads	8
Batch size	32
Hidden	513
Dropout	0.1
Learning rate	0.0001

Table 3. The results of Transformer-based Dungan character-to-unit conversion.

Precision	Recall	F1
90.12	89.91	90.01

3.2. Evaluation on Transfer Learning-Based Dungan Acoustic Modols

3.2.1. Corpus

In the experiment, we utilized recordings of nine female and thirty-one male speakers from the Tsinghua Chinese 30-hour database [45] (totaling 13,389 sentences) as the Mandarin corpus. For the Dungan corpus, we selected five male speakers' recordings (923 per person, totaling 4615 sentences and 6 h). The Dungan corpus encompasses all initial and final pronunciations of the Dungan language. The average sentence length is 18 syllables, with an average duration of 10 s. All recordings were converted to a monochannel 16 kHz sampling frequency with 16-bit quantization accuracy.

3.2.2. Experimental Setup

Three kinds of TTS frameworks, including Tacotron+Griffin-Lim, Tacotron2+WaveNet, and Tacotron2+WaveRNN, were compared in the experiments. Some hyperparameters of the frameworks are provided in Table 4.

Table 4. Model hyperparameters of Tacotron and Tacotron2.

Model		Tacotron	Tacotron2	Forward-Attention Tacotron2
Vocoder		Griffin-Lim	WaveNet	WaveRNN
Encoder	Embedding	Phomeme (256)	Phomeme (512)	Phomeme (512)
	Pre-net	FFN (256, 128)	-	FFN Phomeme (512, 256)
	Encoder core	CBHG (256)	CNN (512) Bi-LSTM (512)	CNN (256) Bi-LSTM (256, 512)
Decoder	Post-net	CBHG (256)	CNN (512)	CNN (512)
	Decoder RNN	GRU (256, 256)	-	LSTM (512, 256)
	Attention	Additive (256)	Location-sensitive (128)	Forward (256)
	Attention RNN	GRU (256)	LSTM (1024, 1024)	LSTM (256)
	Pre-net	FFN (256, 128)	FFN (256, 256)	FFN (256, 128)
Parameter		7.6×10^6	28.9×10^6	23.7×10^6

All three frameworks comprise a front-end text analyzer module, an acoustic model training module, and a vocoder. The text analyzer module transforms Dungan or Chinese sentences into a Pinyin-represented unit sequence, including initials, finals, and their tones and prosodic boundary labels. In the acoustic model training module, we derive the log magnitude spectrogram from the speech signal using Hann windowing with an 80 ms frame length, 12.5 ms frameshift, and a 2048-point Fourier transform.

For the Tacotron+Griffin-Lim framework, acoustic models are trained using an output layer reduction factor of $r = 3$ and the Adam optimizer with a decaying learning rate. The learning rate commences at 0.001 and is subsequently reduced to 0.0005, 0.0003, and 0.0001 after 5, 20, and 50 epochs, respectively. A straightforward loss function is employed for the seq2seq decoder (Mel spectrogram) and the postprocessing network (linear spectrogram). The training batch size is set to 32, with all sequences padding to a maximum length by reconstructing the zero-padded frames. The Griffin-Lim algorithm is utilized as the vocoder for Mel spectrum-to-speech conversion.

For the Tacotron2+WaveNet-based framework, we train the acoustic models using the standard maximum-likelihood training procedure, which involves feeding the correct output instead of the predicted output on the decoder side. This was completed with a batch size of 32. The Adam optimizer was utilized with parameters set as follows: $\beta = 0.9$, $\beta = 0.999$, $\epsilon = 10^{-6}$. The learning rate was initialized at 10^{-3} and then exponentially decayed to 10^{-5} after $50,000$. Additionally, we applied $L2$ regularization with a weight of 10^{-6}. For the Mel spectrum-to-speech conversion, the WaveNet was employed as the vocoder.

In our Tacotron2+WaveRNN-based transfer learning framework, we initially employ a large-scale Mandarin corpus to pre-train a Mandarin acoustic model for subsequent model transfer. This pre-trained model is then used to train the Dungan acoustic model via transfer learning from the Mandarin–Dungan corpus. For vocoding, we utilize the WaveRNN for Mel spectrum-to-speech conversion. Given that parameter settings significantly impact model accuracy and robustness, we optimized these parameters through iterative training and updates.

Each TTS framework implements a monolingual speech synthesis for Mandarin or Dungan and a bilingual one based on transfer learning. We trained several models across three

TTS frameworks to assess the synthesized speech's quality and clarity. In our experiment, 10% of the utterances were randomly allocated to the test set, another 10% were designated for the development set, and the remaining utterances constituted the training set.

Dungan Monolingual Speaker-Dependent Model

We trained the Dungan Monolingual Speaker-Dependent (DSD) acoustic model using recordings from five male speakers, each contributing 923 sentences, totaling 4615 sentences and spanning 6 h. We then compared the quality and clarity of synthesized speech across three frameworks: DSD-Tacotron+Griffin-Lim, DSD Tacotron2+WaveNet, and DSD-Tacotron2+WaveRNN.

Mandarin Monolingual Speaker-Dependent Model

We utilized recordings from nine female and thirty-one male speakers (Tsinghua Chinese 30-h database, consisting of 13,389 sentences) to train the Mandarin Monolingual Speaker-Dependent (MSD) acoustic model. We compared the synthesized speech quality and clarity across three frameworks: MSD-Tacotron+Griffin-Lim, MSD-Tacotron2+WaveNet, and MSD-Tacotron2+WaveRNN.

Mandarin and Dungan Bilingual Speaker-Dependent Model

We utilized recordings from five Dungan male speakers (923 sentences per individual, summing up to 4615 sentences, equivalent to 6 h) as the training data to transfer the Mandarin acoustic model to the Dungan acoustic model to realize a Dungan Speaker-Dependent (MDSD) acoustic model and a Mandarin Speaker-Dependent (MDSM) acoustic model. We then compared the quality and clarity of synthesized speech across six frameworks.

- MDSD-Tacotron+Griffin-Lim
- MDSM-Tacotron+Griffin-Lim
- MDSD-Tacotron2+WaveNet
- MDSM-Tacotron2+WaveNet
- MDSD-Tacotron2+WaveRNN
- MDSM-Tacotron2+WaveRNN

3.2.3. Objective Evaluations

We employed the Mel-cepstral distortion (MCD) [46], Band A Periodicity Distortion (BAP) [47], Root mean squared error (RMSE) [48] and Voiced/Unvoiced error (V/UV) [47] to evaluate the various models objectively. The results for the DSD and MSD acoustic models are presented in Table 5 and Table 6, respectively. Similarly, the MDSM and MDSD results are displayed in Table 7 and Table 8, respectively.

Table 5. Objective results of DSD acoustic model for Dungan.

Model	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	9.675	9.572	9.502
BAP (dB)	0.189	0.187	0.170
F0 RMSE (Hz)	32.785	32.692	32.087
V/UV (%)	9.867	9.721	9.875

Table 6. Objective results of MSD acoustic model for Mandarin.

Model	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	5.460	5.291	5.036
BAP (dB)	0.174	0.171	0.169
F0 RMSE (Hz)	14.629	13.986	13.647
V/UV (%)	5.619	5.793	5.762

Table 7. Objective results of MDSD acoustic model for Dungan.

Model	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	7.523	7.419	7.395
BAP (dB)	0.178	0.175	0.174
F0 RMSE (Hz)	26.891	26.753	26.617
V/UV (%)	7.774	7.693	7.607

Table 8. Objective results of MDSM acoustic model for Mandarin.

Model	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD(dB)	5.339	5.241	5.108
BAP (dB)	0.174	0.173	0.171
F0 RMSE (Hz)	13.775	13.326	13.092
V/UV (%)	5.542	5.472	5.481

In the context of low-resource Dungan speech synthesis, the quality of attention alignment between the encoder and decoder significantly influences the quality of synthesized speech. Misalignments are primarily evident in readability, skipping, and repetition. Consequently, we employ the Diagonal Focus Rate (DFR) and Word-Level Intelligibility Rate (IR) [49] to assess readability in low-resource languages, as illustrated in Table 9. The DFR represents the attention map between the encoder and decoder, serving as an architectural metric. The IR measures the percentage of test words pronounced correctly and clearly by humans, a standard metric for assessing the quality of low-resource speech generation.

Table 9. Readability of synthesized Dungan speech.

Model	IR (%)	DFR (%)
DSD-Tacotron+Griffin-Lim	82.93	79.64
DSD-Tacotron2+WaveNet	86.67	82.43
DSD-Tacotron2+WaveRNN	89.41	84.39
MDSD-Tacotron+Griffin-Lim	95.03	91.14
MDSD-Tacotron2+WaveNet	96.69	94.43
MDSD-Tacotron2+WaveRNN	98.47	97.39

3.2.4. Subjective Evaluation

For subjective evaluations, 30 sentences were randomly selected from the test set. We conducted three tests: mean opinion score (MOS), degradation mean opinion score (DMOS), and AB preference to assess the quality of synthesized speech. We recruited 20 native Mandarin speakers and 10 native Dungan international students (who understood Chinese) as participants. These participants received training before the formal evaluation. Mandarin participants evaluated the Mandarin acoustic models of MSD and MDSM, whereas Dungan participants assessed the Dungan acoustic models of DSD and MDSD. During the MOS test, participants rated the naturalness of synthesized speech on a 5-point scale. The average MOS scores for synthesized Dungan and Mandarin speech are presented in Figures 7 and 8.

In the DMOS test, each model's synthesized utterance and corresponding original recording comprised a pair of speech files. These pairs were randomly played to the subjects, with the synthesized speech preceding the original. The participants were tasked with meticulously comparing the two files and rating the similarity of the synthesized speech to the original on a 5-point scale. A score of 5 indicated that the synthesized speech was similar to the original, whereas a score of 1 signified a significant disparity. Figures 9 and 10 display the average DMOS scores for synthesized Dungan and Mandarin speech, respectively.

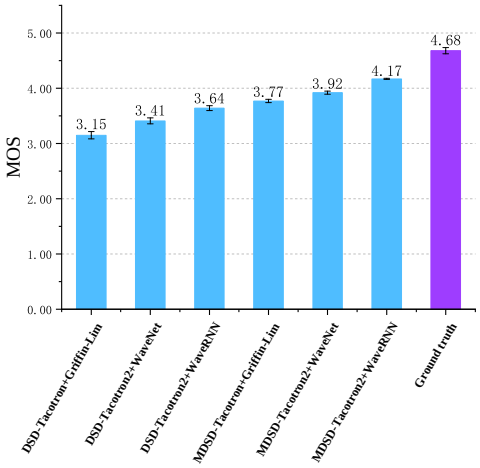

Figure 7. The average MOS scores of synthesized Dungan speech under 95% confidence intervals.

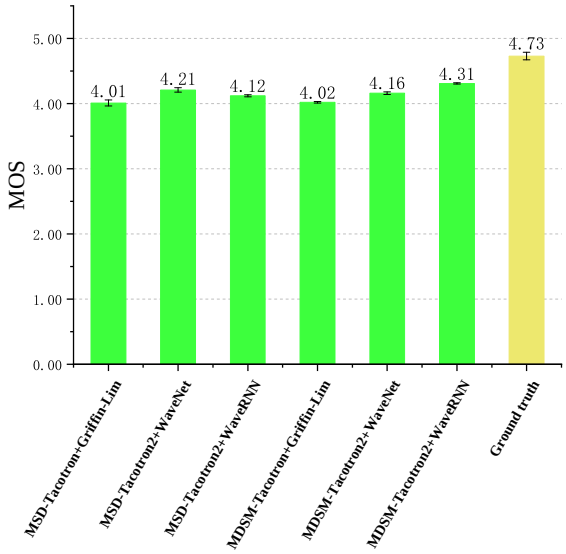

Figure 8. The average MOS scores of synthesized Mandarin speech under 95% confidence intervals.

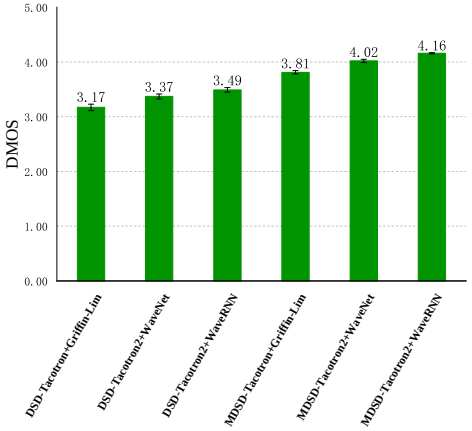

Figure 9. The average DMOS scores of synthesized Dungan speech under 95% confidence intervals.

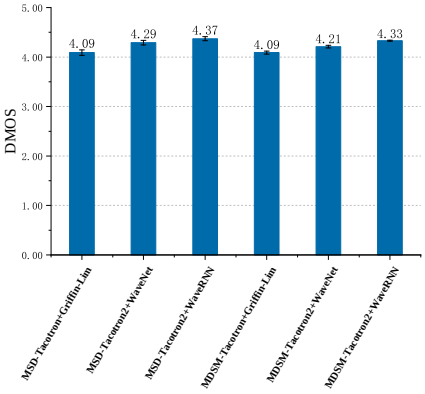

Figure 10. The average DMOS scores of synthesized Mandarin speech under 95% confidence intervals.

In the AB preference test, each pair consisted of two identical sentences. The synthesized utterances were played in a randomized order. Participants were instructed to listen and evaluate which utterance had superior quality or indicate "neutral" if no preference was discerned. The synthesized Dungan and Mandarin speech preference outcomes are presented in Tables 10 and Tables 11, respectively.

Table 10. Subjective AB preference score (%) of Dungan with $\rho < 0.01$.

	DSD-Tacotron+ Griffin-Lim	DSD-Tacotron2+ WaveNet	DSD-Tacotron2 +WaveRNN	MDSD-Tacotron+ Griffin-Lim	MDSD-Tacotron2 +WaveNet	MDSD-Tacotron2+ WaveRNN	Neutral
1	12.7	22.9	52.6	-	-	-	11.8
2	29.5	32.0	27.6	-	-	-	10.9
3	-	-	-	17.7	-	69.9	12.4
4	-	-	-	3.2		70.8	11.3
5	-	-	-	-	17.1	72.1	10.8

Table 11. Subjective AB preference score(%) of Mandarin with $\rho < 0.01$.

	MSD-Tacotron+ Griffin-Lim	MSD-Tacotron2+ WaveNet	MSD-Tacotron2+ WaveRNN	MDSM-Tacotron+ Griffin-Lim	MDSM-Tacotron2+ WaveNet	MDSM-Tacotron2+ WaveRNN	Neutral
1	-	24.54	63.56	-	-	-	11.9
2	-	19.98	67.42	-	-	-	12.6
3	-	-	-	-	11.8	71.9	16.3
4	-	-	-	14.4	-	75.1	10.5
5	-	-	-	-	10.7	79.6	9.7

4. Discussion

In objective evaluations, although the Tacotron+Griffin-Lim-based TTS framework maps linguistic features to acoustic features frame by frame through the monolingual Dungan corpus, the synthesized Dungan speech needs to improve its quality and readability. However, the forward attention and fine-tuned acoustic model can enhance readability and reduce training time. Consequently, the transfer learning-based Tacotron2+WaveRNN framework's acoustic model outperforms others. The objective results of the MDSD acoustic model surpass those of the DSD acoustic model. This is because Dungan is a variation of China's Northwestern dialect, which shares many internal similarities. Given the pronunciation similarities between Mandarin and Dungan, the same symbol represents their exact pronunciations. Therefore, we conclude that adding a Mandarin corpus and using transfer learning can improve the quality and readability of synthesized Dungan speech.

All subjective evaluations align with objective assessments in various aspects. The transfer learning-based Tacotron2+waveRNN framework yields superior speech quality,

particularly regarding the naturalness and readability of synthesized speech. With the addition of the Mandarin corpus, the quality and readability of synthesized Dungan speech using the transfer learning-based TTS frameworks surpass those monolingual corpus-trained TTS frameworks. This is further validated by the AB preference test, which confirms that our proposed TTS frameworks offer improved quality and readability compared to the speech synthesized by the monolingual acoustic model.

5. Conclusions

This study extends our prior research by implementing a transfer learning-based Mandarin speech synthesis and a low-resource Dungan speech synthesis under the Tacotron2+WaveRNN framework. We also developed a comprehensive Dungan text analyzer. Objective and subjective experiments revealed that the transfer learning-based Dungan speech synthesis under the Tacotron2+WaveRNN framework outperformed alternative methods and the monolingual Dungan speech synthesis framework. Furthermore, transfer learning did not compromise the speech quality and readability of the synthesized low-resource Dungan speech. Therefore, our approach holds significant potential for developing speech synthesis systems for low-resource minority languages.

Numerous breakthroughs have been achieved in TTS based on deep neural networks. We have noticed that some new speech synthesis methods [50–52] have been proposed recently. Motivated by recent advancements in auto-regressive (AR) models employing decoder-only architectures for text generation, several studies, such as VALL-E [53] and BASE TTS [54], apply similar architectures to TTS tasks. These studies demonstrate the remarkable capacity of decoder-only architectures to produce natural-sounding speech. These studies demonstrate the remarkable capacity of decoder-only architectures to produce natural-sounding speech. Future research will focus on using these new methods to improve the quality of Dungan language speech synthesis, reduce the Dungan corpus size, and achieve speech synthesis for Dungan languages using a larger corpus. Additionally, multitask learning will be explored to realize speaker-independent scenarios and enhance the emotion of synthesized Dungan speech.

Author Contributions: Conceptualization, M.L. and H.Y.; formal analysis, H.Y. and R.J.; data curation, M.L. and R.J.; writing—original draft preparation, M.L. and R.J.; writing—review and editing, H.Y. and M.L.; supervision, H.Y.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the research fund from the National Natural Science Foundation of China (Grant No. 62067008).

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used two training datasets in the manuscript. One is a publicly available Mandarin dataset (THCHS-30), and the other is a Donggan dataset, including speech and text. The former has been public and can be accessed from http://www.openslr.org/18/ (accessed on 16 June 2024). The latter is a self-constructed dataset and is not publicly available. However, the data will be made available on request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Tu, T.; Chen, Y.J.; Chieh Yeh, C.; Yi Lee, H. End-to-end Text-to-speech for Low-resource Languages by Cross-Lingual Transfer Learning. *arXiv* **2019**, arXiv:1904.06508.
2. Liu, R.; Sisman, B.; Bao, F.; Yang, J.; Gao, G.; Li, H. Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 274–285. [CrossRef]
3. Saeki, T.; Maiti, S.; Li, X.; Watanabe, S.; Takamichi, S.; Saruwatari, H. Text-Inductive Graphone-Based Language Adaptation for Low-Resource Speech Synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1829–1844. [CrossRef]

4. Xu, J.; Tan, X.; Ren, Y.; Qin, T.; Li, J.; Zhao, S.; Liu, T.Y. LRSpeech: Extremely Low-Resource Speech Synthesis and Recognition. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'20, New York, NY, USA, 6–10 July 2020; pp. 2802–2812. [CrossRef]
5. He, M.; Yang, J.; He, L.; Soong, F.K. Multilingual Byte2Speech Models for Scalable Low-resource Speech Synthesis. *arXiv* **2021**, arXiv:2103.03541.
6. Oliveira, F.S.; Casanova, E.; Junior, A.C.; Soares, A.S.; Galvão Filho, A.R. CML-TTS: A Multilingual Dataset for Speech Synthesis in Low-Resource Languages. In *Text, Speech, and Dialogue*; Ekštein, K., Pártl, F., Konopík, M., Eds.; Springer: Cham, Switzerland, 2023; pp. 188–199.
7. Zhu, Y. Donggan Language: A Special Variety of the Shaanxi and Gansu Dialects. *Asian Lang. Cult.* **2013**, *4*, 51–60.
8. Jiang, Y. Donggan Language and Its Relation to the Shaanxi and Gansu Dialects. *J. Chin. Linguist.* **2014**, *42*, 229–258.
9. Chen, L.; Yang, H.; Wang, H. Research on Dungan speech synthesis based on Deep Neural Network. In Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), Taipei, Taiwan, 26–29 November 2018; pp. 46–50. [CrossRef]
10. Jiang, R.; Chen, C.; Shan, X.; Yang, H. Using Speech Enhancement to Realize Speech Synthesis of Low-Resource Dungan Languages. In Proceedings of the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Singapore, 18–20 November 2021; pp. 193–198. [CrossRef]
11. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 1, pp. 373–376.
12. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Speech synthesis based on hidden Markov models. *Proc. IEEE* **2013**, *101*, 1234–1252. [CrossRef]
13. Ling, Z.H.; Deng, L.; Yu, D. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 2129–2139. [CrossRef]
14. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966. [CrossRef]
15. Wang, P.; Qian, Y.; Soong, F.K.; He, L.; Zhao, H. Word embedding for recurrent neural network based TTS synthesis. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 4879–4883. [CrossRef]
16. Yu, Q.; Liu, P.; Wu, Z.; Ang, S.K.; Meng, H.; Cai, L. Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5545–5549. [CrossRef]
17. Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4234–4245. [CrossRef]
18. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the 18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, 20–24 August 2017.
19. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783. [CrossRef]
20. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
21. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499.
22. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Van den Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. *arXiv* **2018**, arXiv:1802.08435.
23. Byambadorj, Z.; Nishimura, R.; Ayush, A.; Ohta, K.; Kitaoka, N. Text-to-speech system for low-resource language using cross-lingual transfer learning and data augmentation. *EURASIP J. Audio Speech Music. Process.* **2021**, *2021*, 42. [CrossRef]
24. Joshi, R.; Garera, N. Rapid Speaker Adaptation in Low Resource Text to Speech Systems using Synthetic Data and Transfer learning. In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, Hong Kong, China, 2–4 December 2023; Huang, C.R., Harada, Y., Kim, J.B., Chen, S., Hsu, Y.Y., Chersoni, E.A.P., Zeng, W.H., Peng, B., Li, Y., et al., Eds.; ACL: Hong Kong, China, 2023; pp. 267–273.
25. Do, P.; Coler, M.; Dijkstra, J.; Klabbers, E. Strategies in Transfer Learning for Low-Resource Speech Synthesis: Phone Mapping, Features Input, and Source Language Selection. In Proceedings of the 12th ISCA Speech Synthesis Workshop (SSW2023), Grenoble, France, 26–28 August 2023; pp. 21–26. [CrossRef]

26. Azizah, K.; Jatmiko, W. Transfer learning, style control, and speaker reconstruction loss for zero-shot multilingual multi-speaker text-to-speech on low-resource languages. *IEEE Access* **2022**, *10*, 5895–5911. [CrossRef]
27. Cai, Z.; Yang, Y.; Li, M. Cross-lingual multi-speaker speech synthesis with limited bilingual training data. *Comput. Speech Lang.* **2023**, *77*, 101427. [CrossRef]
28. Yang, H.; Oura, K.; Wang, H.; Gan, Z.; Tokuda, K. Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis. *Multimed. Tools Appl.* **2015**, *74*, 9927–9942. [CrossRef]
29. Wang, L.; Yang, H. Tibetan word segmentation method based on bilstm_ crf model. In Proceedings of the IEEE 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia, 15–17 November 2018; pp. 297–302.
30. Zhang, W.; Yang, H.; Bu, X.; Wang, L. Deep learning for mandarin-tibetan cross-lingual speech synthesis. *IEEE Access* **2019**, *7*, 167884–167894. [CrossRef]
31. Zhang, W.; Yang, H. Improving Sequence-to-sequence Tibetan Speech Synthesis with Prosodic Information. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 6012. [CrossRef]
32. Zhang, W.; Yang, H. Meta-Learning for Mandarin-Tibetan Cross-Lingual Speech Synthesis. *Appl. Sci.* **2022**, *12*, 2185. [CrossRef]
33. Hai, F. A Pilot Study of Loan Words in Central-Asian Dungan Language. *Xinjiang Univ. J.* **2000**, *28*, 58–63.
34. Lin, T. Features, Situation and Development Trends of Tung'gan Language in Central Asia. *Contemp. Linguist.* **2016**, *18*, 234–243.
35. Gladney, D.C. Relational alterity: Constructing dungan (hui), uygur, and Kazakh identities across China, central Asia, and Turkey. *Hist. Anthropol.* **1996**, *9*, 445–477. [CrossRef]
36. Miao, D.X. Bilingual Teaching Model of the Donggan People. *J. Res. Educ. Ethn. Minor.* **2008**, *19*, 111–114.
37. Jia, Y.; Huang, D.; Liu, W.; Dong, Y.; Yu, S.; Wang, H. Text normalization in mandarin text-to-speech system. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 4693–4696. [CrossRef]
38. Wanmezhaxi, N. Research on Several Key Issues in Tibetan Word Segmentation. *J. Chin. Inf. Process.* **2014**, *28*, 132–139.
39. Zavyalova, O. Dungan Language. 2015. Available online: https://www.academia.edu/42869092/Dungan_Language (accessed on 16 June 2024).
40. Lin, T. Donggan Writing—A Successful Trial of Chinese Alphabetic Writing. *J. Second. Northwest Univ. Natl.* **2005**, *2005*, 31–36.
41. Yang, W.J.; Zhang, R. Ethnic Identity in the Context of Across-nation—A Study Case of "Dunggan" and the Hui Nationality. *J. South-Cent. Univ. Natl.* **2009**, *29*, 31–36.
42. Zheng, Y.; Tao, J.; Wen, Z.; Li, Y. BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End. *Proc. Interspeech* **2018**, *9*, 47–51. [CrossRef]
43. Hlaing, A.M.; Pa, W.P. Sequence-to-Sequence Models for Grapheme to Phoneme Conversion on Large Myanmar Pronunciation Dictionary. In Proceedings of the 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Cebu, Philippines, 25–27 October 2019; pp. 1–5. [CrossRef]
44. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. In *Artificial Neural Networks and Machine Learning—ICANN 2018*; Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I., Eds.; Springer: Cham, Switzerland, 2018; pp. 270–279.
45. Wang, D.; Zhang, X. THCHS-30: A Free Chinese Speech Corpus. *arXiv* **2015**, arXiv:1512.01882.
46. Kubichek, R. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of the IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Canada, 19–21 May 1993; Volume 1, pp. 125–128. [CrossRef]
47. Dhiman, J.K.; Seelamantula, C.S. A Spectro-temporal Technique for Estimating Aperiodicity and Voiced/unvoiced Decision Boundaries of Speech Signals. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019), Brighton, UK, 12–17 May 2019; pp. 6510–6514. [CrossRef]
48. Castelazo, I.; Mitani, Y. On the use of the mean squared error as a proficiency index. *Accredit. Qual. Assur.* **2012**, *17*, 95–97. [CrossRef]
49. Ren, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Almost Unsupervised Text to Speech and Automatic Speech Recognition. *arXiv* **2020**, arXiv:1905.06791.
50. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. *arXiv* **2022**, arXiv:2006.04558.
51. Chen, J.; Song, X.; Peng, Z.; Zhang, B.; Pan, F.; Wu, Z. LightGrad: Lightweight Diffusion Probabilistic Model for Text-to-Speech. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2023), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
52. Guo, Y.; Du, C.; Ma, Z.; Chen, X.; Yu, K. VoiceFlow: Efficient Text-To-Speech with Rectified Flow Matching. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2024), Seoul, Republic of Korea, 14–19 April 2024; pp. 11121–11125. [CrossRef]

53. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv* **2023**, arXiv:2301.02111.
54. Łajszczak, M.; Cámbara, G.; Li, Y.; Beyhan, F.; van Korlaar, A.; Yang, F.; Joly, A.; Martín-Cortinas, Á.; Abbas, A.; Michalski, A.; et al. BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data. *arXiv* **2024**, arXiv:2402.08093.