angewandte Wissenschaften

MDPI

Artikel

Mit Transferlernen Dungan mit geringen Ressourcen realisieren Sprache Sprachsynthese

Mengrui Liu 1,†, Rui Jiang 2,† und Hongwu Yang 2,3,*

[1] Hochschule für Elektro- und Informationstechnik, Tongji-Universität, Shanghai 201804, China; liuxh709@163.com

[2] Schule für Bildungstechnologie, Northwest Normal University, Lanzhou 730070, China; jiangh940618@163.com

[3] Schlüssellabor für Bildungsdigitalisierung der Provinz Gansu, Lanzhou 730070, China

* Korrespondenz: yanghw@nwnu.edu.cn

† Diese Autoren haben zu gleichen Teilen zu der Arbeit beigetragen.

Zusammenfassung: Dieser Artikel präsentiert eine auf Transferlernen basierende Methode zur Verbesserung der Qualität der synthetisierten Sprache der ressourcenarmen Dungan-Sprache. Diese Verbesserung wird durch die Feinabstimmung eines vorab trainierten Mandarin-Akustikmodells zu einem Dungan-Sprachakustikmodell unter Verwendung eines begrenzten Dungan-Korpus innerhalb des Tacotron2+WaveRNN-Frameworks erreicht. Unsere Methode beginnt mit der Entwicklung eines transformerbasierten Dungan-Textanalysators, der in der Lage ist, Einheitssequenzen mit eingebetteten prosodischen Informationen aus Dungan-Sätzen zu generieren. Diese Einheitssequenzen liefern zusammen mit den Sprachmerkmalen <Einheitssequenz mit prosodischen Bezeichnungen, Mel-Spektrogramme>-Paare als Eingabe von Tacotron2 zum Trainieren des Akustikmodells. Gleichzeitig haben wir ein Tacotron2-basiertes Mandarin-Akustikmodell unter Verwendung eines groß angelegten Mandarin-Korpus vorab trainiert. Das Modell wird dann mit einem klein angelegten Dungan-Sprachkorpus feinabgestimmt, um ein Dungan-Akustikmodell abzuleiten, das die Ausrichtung und Zuordnung der Einheiten zu den Spektrogrammen autonom lernt . Die resultierenden Spektrogramme werden über den WaveRNN-Vocoder in Wellenformen umgewandelt, was die Synthese hochwertiger Mandarin- oder Dungan-Sprache erleichtert. Sowohl subjektive als auch objektive Experimente deuten darauf hin, dass die vorgeschlagene, auf Transferlernen basierende Dungan- Sprachsynthese im Vergleich zu Modellen, die nur mit dem Dungan-Korpus und anderen Methoden trainiert wurden, bessere Ergebnisse erzielt. Folglich bietet unsere Methode eine Strategie zur Erzielung einer Sprachsynthese für ressourcenarme Sprachen, indem prosodische Informationen hinzugefügt und ein ähnliches, ressourcenreiches Sprachkorpus durch Transferlernen genutzt wird.

Schlüsselwörter: Sprachsynthese der Dungan-Sprache; Textanalyse; Transferlernen; ressourcenarme Sprache; Tacotron2

Zitat: Liu, M.; Jiang, R.; Yang, H. Einsatz von Transferlernen zur Realisierung einer ressourcenarmen Sprachsynthese in der Dungan-Sprache. Appl. Sci. **2024,** 14, 6336. https://doi.org/10.3390/app14146336

Wissenschaftliche Herausgeber: Gloria Corpas Pastor und Tharindu Ranasinghe

Empfangen: 17. Juni 2024
Überarbeitet: 17. Juli 2024
Akzeptiert: 18. Juli 2024
Veröffentlicht: 20. Juli 2024

1. Einleitung

Sprachsynthese (Text-to-Speech (TTS)-Konvertierung) wird häufig in Smart Homes, Navigationssystemen und Hörbuchanwendungen verwendet. Weltweit gibt es etwa 6.000 Sprachen , von denen die meisten als ressourcenarm gelten. Während bei der Sprachsynthese wichtiger Sprachen wie Mandarin, Englisch und Französisch erhebliche Fortschritte erzielt wurden, ist die Sprachqualität von TTS für ressourcenarme Sprachen wie Tibetisch und Dunganisch nach wie vor nicht optimal. In den letzten Jahren gab es einen Anstieg der Forschung zur Sprachsynthese für ressourcenarme Sprachen, wie zahlreiche Studien belegen [1–6]. Die Forschung zur Sprachsynthese der Dunganisch-Sprache muss jedoch noch abgeschlossen werden. Die Dunganisch-Sprache, eine Variante der Shanxi-Gansu-Dialekte innerhalb des in Zentralasien gesprochenen chinesischen Dialekts, wird aufgrund ihrer eingeschränkten Verwendung, der sinkenden Zahl von Sprechern und des Mangels an sprachlichem Material als ressourcenarme Sprache eingestuft [7,8]. Da Russisch zur Amtssprache Zentralasiens geworden ist , stellt die Erstellung eines umfassenden Sprachkorpus mit linguistischem Wissen für eine qualitativ hochwertige Dungan-Sprachsynthese eine große Herausforderun

DNN-basierte Dungan-Sprachsynthese [9,10], die Qualität der synthetisierten Sprache war aufgrund des begrenzten Trainingskorpus nicht hoch.

Zu den Sprachsynthesetechnologien gehören die konkatenative Sprachsynthese auf Basis von Einheitenauswahl [11], die statistische parametrische Sprachsynthese (SPSS) auf Basis von Hidden-Markov-Modellen (HMM) [12] und die Deep-Learning-basierte Sprachsynthese [13,14]. Während Deep Learning die Sprachsynthesetechnologie erheblich weiterentwickelt hat, haben Methoden wie Long Short-Term Memory (LSTM) und bidirektionales LSTM [15,16] zeitliche Informationsbeschränkungen behoben. Darüber hinaus haben End-to-End-Sprachsynthesemodelle [17] wie Tacotron [18] und Tacotron2 [19] die Fähigkeit gezeigt, Text direkt in Sprache umzuwandeln. Wenn diese Modelle mit großen Text-zu-Sprache-Paaren trainiert werden, produzieren sie synthetisierte Sprache unter Verwendung hochwertiger Vocoder wie dem Griffin-Lim-Algorithmus [20], WaveNet [21] und WaveRNN [22].

Solche Systeme erfordern jedoch umfangreiche Trainingskorpora. Bei ressourcenarmen Sprachen erschwert das Fehlen eines Trainingskorpus es End-to-End-Modellen, die prosodische Struktur von Sätzen zu erlernen. Dies führt zu einem Mangel an prosodischen Änderungen in der synthetisierten Sprache, was ihre Natürlichkeit beeinträchtigt und die Sprachsynthese ressourcenarmer Sprachen vor Herausforderungen stellt.

Um das Problem unzureichender Trainingskorpora für die Sprachsynthese in ressourcenarmen Sprachen zu mildern, wurde sprachübergreifendes Transferlernen [23–25] eingesetzt. Bei dieser Technik wird ein Sprachmodell mit einer Kombination aus einem großen Korpus einer ressourcenreichen Sprache und einem kleineren Korpus einer ressourcenarmen Sprache trainiert, gefolgt von der Anpassung dieses Modells an die ressourcenarme Sprache. Transferlernen in der Sprachsynthese hat sich als effektive Strategie zur Sprachproduktion in ressourcenarmen Sprachen erwiesen, indem die Fähigkeiten eines ressourcenreichen akustischen Sprachmodells genutzt werden [26,27].

In unserer früheren Forschung zur tibetischen Sprachsynthese [28–32] haben wir festgestellt, dass die Integration prosodischer Informationen durch auf Transferlernen basierende Techniken die Qualität der synthetisierten Sprache für ressourcenarme Sprachen wie Tibetisch verbessert. Aufbauend auf dieser Erkenntnis implementiert die vorliegende Studie einen Sequenz-zu-Sequenz-Ansatz (seq2seq) für die Sprachsynthese der Dungan-Sprache, der Transferlernen und prosodische Informationen im Rahmen von Tacotron2+WaveRNN nutzt. Diese Methode beinhaltet die Verwendung eines Dungan- Textanalysators, um prosodische Bezeichnungen aus Dungan-Sätzen für die Modellintegration zu extrahieren, wobei ein auf Tacotron2 basierendes Mandarin-Akustikmodell verwendet und das Dungan-Akustikmodell mit einem begrenzten Dungan-Sprachkorpus feinabgestimmt wird. Die wichtigsten Beiträge werden im Folgenden beschrieben:

- Frontend: Wir haben einen vollständigen Textanalysator für die Dungan-Sprache implementiert, der Module für Textnormalisierung, Wortsegmentierung, Vorhersage prosodischer Grenzen und Einheitengenerierung auf Basis der Transformer-Technologie umfasst. Dieser Analysator kann Initialen und Endungen als Sprachsyntheseeinheiten mit prosodischen Labels aus Dungan-Sätzen erzeugen. • Backend: Wir haben eine seq2seq-Sprachsynthese für die Dungan-Sprache erreicht,

indem wir ein vorab trainiertes Mandarin-Akustikmodell im Rahmen von Tacotron2+WaveRNN angepasst haben.

Dies wurde erreicht, indem die ortsabhängige Aufmerksamkeit von Tacotron2 durch eine Vorwärtsaufmerksamkeit ersetzt wurde, wodurch die Konvergenzgeschwindigkeit und Stabilität verbessert wurde.

Der Rest des Artikels ist wie folgt aufgebaut. In Abschnitt 2 stellen wir zunächst unser auf Transferlernen basierendes Dungan-Sprachsynthese-Framework unter Tacotron2+WaveRNN vor. Der Versuchsaufbau und die Ergebnisse werden in Abschnitt 3 vorgestellt, während die Ergebnisse in Abschnitt 4 diskutiert werden. Abschließend werden in Abschnitt 5 eine kurze Schlussfolgerung und ein Überblick für zukünftige Arbeiten gegeben.

2. Modelle und Methoden

Das vorgeschlagene Framework für die ressourcenarme Dungan-Sprachsynthese auf Basis von Transferlernen ist in Abbildung 1 dargestellt und umfasst ein Modul zur Merkmalsextraktion, ein vorab trainiertes Mandarin-Akustikmodell, ein Trainingsmodul für ein Dungan-Akustikmodell auf Basis von Transferlernen und einen auf dem WaveRNN-Vocoder basierenden Sprachsynthesizer.

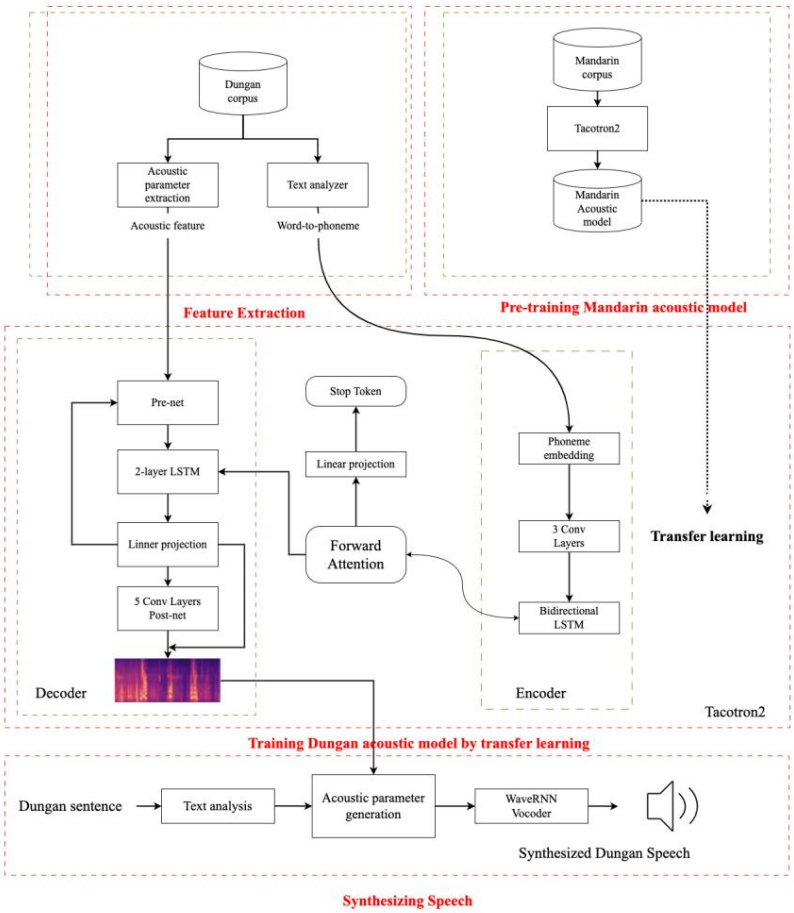

Abbildung 1. Das Framework der auf Tacotron2+WaveRNN basierenden Dungan-Sprachsynthese.

Das Merkmalsextraktionsmodul extrahiert akustische Merkmale wie das Mel-Spektrogramm aus Sprachsignalen und Sequenzen von Sprachsyntheseeinheiten aus Sätzen. Wir haben einen vollständigen Textanalysator für die Dungan-Sprache entwickelt, um Sprachsyntheseeinheiten mit prosodischen Merkmalen zu extrahieren und Dungan-Sätze auf Einheitensequenzen abzubilden. Da sowohl Mandarin als auch Dungan Anlaute und Endlaute als zentrale Sprachsyntheseeinheiten verwenden, enthält die resultierende Einheitensequenz diese Elemente und relevante prosodische Informationen, einschließlich Silbentönen und prosodischen Grenzbezeichnungen auf Satzebene.

Da Tacotron2 eines der beliebtesten Encoder-to-Decoder-Frameworks für die Sprachsynthese ist und der WaveRNN-Vocoder natürliche Sprache erzeugen kann, verwenden wir Tacotron2 zum Trainieren akustischer Modelle und WAVRNN zum Konvertieren von Spektrogrammen in Wellenformen sowohl für die Dungan-Sprache als auch für Mandarin. Das Mandarin-Akustikmodell ist mit einem umfangreichen Mandarin-Korpus vortrainiert, während das Dungan-Sprachmodell mit einem kleinen Dungan-Korpus vom Mandarin-Akustikmodell übertragen wird.

In der Phase der Sprachsynthese generiert der WaveRNN-Vocoder Dungan- oder Mandarin- Sprache aus den Eingaben von Dungan- oder chinesischen Sätzen. Der Textanalysator generiert zunächst die kontextabhängigen Bezeichnungen aus dem Eingabesatz. Anschließend werden die Sprachsynthese-Einheitssequenzen (Anfänge und Endungen mit ihren prosodischen Informationen) in das Mandarin- oder Dungan-Akustikmodell eingespeist, um das Mel-Spektrogramm zu generieren. Der WaveRNN-Vocoder wird schließlich verwendet, um die Sprachwellenformen aus dem Mel-Spektrogramm zu generieren. Für die Analyse chinesischer Texte verwenden wir einen selbst entwickelten chinesischen Textanalysator.

2.1. Textanalysator der Dungan-Sprache

Im Gegensatz zu den vorherrschenden seq2seq-Sprachsynthesetechniken, die für die wichtigsten Sprachen entwickelt wurden und ausschließlich das Paar <Phonemsequenz, Sprache> zum Trainieren akustischer Modelle verwenden, verwendet unser Ansatz eine Einheitssequenz, die prosodische Bezeichnungen enthält, wi

Ton jeder Silbe und die prosodische Grenze eines Satzes, die als „Phonemfolge“ dient. Folglich ist es wichtig, einen umfassenden Textanalysator zu entwickeln, der die Einheitssequenzen eines Satzes und ihre prosodischen Bezeichnungen extrahieren kann. Zu diesem Zweck haben wir mithilfe unseres hauseigenen chinesischen Textanalysators einen Textanalysator für die Dungan-Sprache entwickelt , wie in Abbildung 2 dargestellt. Der Prozess beginnt mit der Normalisierung und Segmentierung des eingegebenen Dungan-Satzes, um die Wortgrenze zu bestimmen. Darauf folgt eine prosodische Grenzanalyse, um sowohl die prosodische Wort- als auch die prosodische Phrasengrenze zu identifizieren. Im letzten Schritt werden die Initialen und Endungen der Dungan-Zeichen durch einen transformatorbasierten Zeichen-zu-Einheit-Konvertierungsprozess abgeleitet.

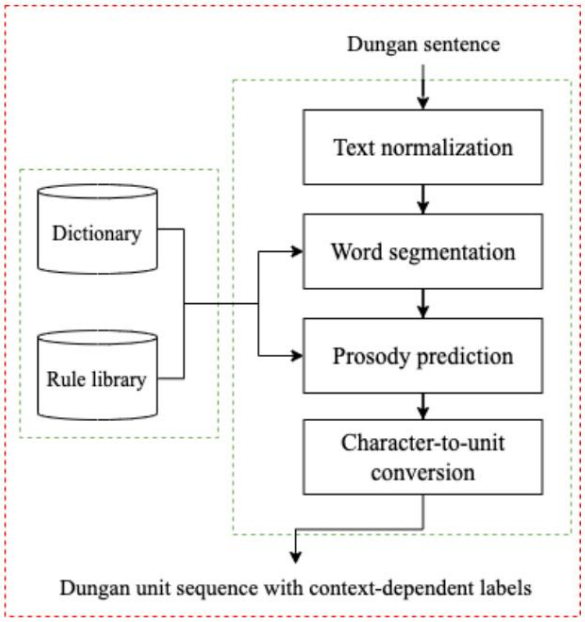

Abbildung 2. Verfahren zur Dungan-Textanalyse.

2.1.1. Sprachsyntheseeinheit der Dungan-Sprache

Obwohl Dunganisch ein anderes Schriftsystem verwendet, ist es außerhalb Chinas eine dialektale Aussprache des Mandarin. Die Dunganisch-Sprache wird in kyrillischer Schrift geschrieben und ähnelt slawischen Sprachen wie Russisch. Die Dunganisch-Sprache besteht also aus phonetischen Zeichen mit sequenzieller Schreibweise, die einer dem Chinesischen ähnlichen Struktur folgen [33–35]. Die Schreibweise der Dunganisch-Zeichen besteht aus Initialen, Endungen und Ton, wie in Abbildung 3 dargestellt .

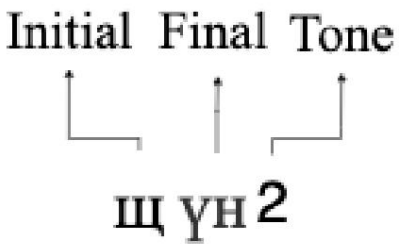

Abbildung 3. Struktur eines Dungan-Charakters.

Dieser Artikel verwendet Initialen und Endlaute als Sprachsyntheseeinheit. Das Dungan-Zeichen besteht aus 25 Initialen (einschließlich der Nullinitiale) und 32 Endlauten, wie in Tabelle 1 gezeigt.
Wie im Mandarin sind die Töne der Dungan-Sprache entscheidend für die Unterscheidung von Semantik und Emotionen [36]. Dunganisch verfügt über vier Töne, den hellen Ton ausgenommen, nämlich den gleichmäßigen Ton (21), den steigenden Ton (24), den fallend-steigenden Ton (53) und den fallenden Ton (44), die jeweils durch die Zahlen 1 bis 4 gekennzeichnet sind.

Tabelle 1. Die Initialen und Endungen der Dungan-Sprache.

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

2.1.2. Textnormalisierung

Jeder Eingabesatz kann numerische Formen von Zeit, Datum, Abkürzungen und spezielle geschützte Substantive enthalten. Bevor ein Satz in eine Folge phonetischer Symbole umgewandelt wird, ist es wichtig, eine Textnormalisierung zu verwenden, um nicht standardisierten Text in ein einheitliches phonetisches Symbol umzuwandeln. Daher haben wir eine regelbasierte Textnormalisierung implementiert, um nicht-dunganische Zeichen zu identifizieren. Wir haben einen Satz von dunganischen Textnormalisierungsregeln entwickelt, die auf chinesischen Textnormalisierungsregeln basieren [37] , und haben die Add-Restore-Methode verwendet, um die dunganischen Zeichen gemäß [38] zu normalisieren.

2.1.3. Wortsegmentierung

Wortgrenzen spielen eine wichtige Rolle bei der Vorhersage prosodischer Grenzen. Daher ist es wichtig, die Wortgrenzen eines Satzes nach der Normalisierung zu identifizieren. Dungan-Sätze weisen klare Unterscheidungen zwischen Wörtern und Silben auf, was die Segmentierung relativ unkompliziert macht. Wir haben einen auf maximaler Übereinstimmung basierenden Wortsegmentierungsalgorithmus verwendet, um Dungan-Wörter aus dem Eingabesatz zu extrahieren. Um diesen Prozess zu erleichtern, haben wir ein Dungan-Wortwörterbuch mit 49.293 Wörtern zusammengestellt. Das längste Wort in diesem Wörterbuch umfasst acht Zeichen, während das kürzeste ein einzelnes Zeichen ist. Das Wörterbuch umfasst in erster Linie zentrale Dungan-Begriffe, wie sie in Quellen wie „Common Dictionary of Dungan Language“ [39], „A Survey on Tungan Language in Central Asia“ [40], „A Survey of Dungan Language“ [41] und weiteren durchsuchbaren Dungan-Begriffen, die online verfügbar sind, referenziert werden.

2.1.4. Vorhersage prosodischer Grenzen

Unser Ansatz verwendet Initialen und Endungen zusammen mit ihren prosodischen Bezeichnungen als Eingangssequenz für das akustische Modell. Daher ist das Extrahieren der prosodischen Struktur aus Dungan-Sätzen entscheidend für die Synthese qualitativ hochwertiger Sprache. Wie Mandarin kann Dungans prosodische Hierarchie in prosodische Wörter, prosodische Phrasen, Intonationsphrasen und Satzpausen segmentiert werden. Die Grenzen von Intonationsphrasen können leicht anhand von Dungan-Satzzeichen identifiziert werden. In dieser Studie verwendeten wir ein BiLSTM mit einer auf bedingten Zufallsfeldern (BiLSTM_CRF) basierenden Methode, wie in Abbildung 4 dargestellt , um die Grenzen prosodischer Wörter und Phrasen vorherzusagen [42].

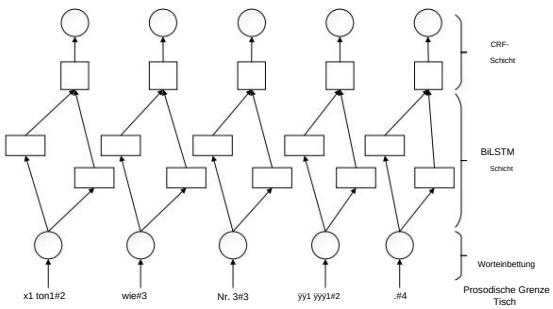

Abbildung 4. Das Framework der BLSTM_CRF-basierten Dungan Prosodic Boundary Prediction. Die Eingabe ist ein Dungan-Satz mit prosodischen Informationen.

Wir verwendeten vier verschiedene prosodische Wortpositions-Beschriftungssätze (#1, #2, #3, #4), um Dungan-Wörter in prosodische Phrasen zu kategorisieren. Konkret wurde #1 verwendet, um die prosodischen Wörter zu kennzeichnen, #2 bezeichnete die prosodischen Phrasen, #3 markierte das Ende eines Dungan-Wortes.

Wort und Nr. 4 bezeichnete eine Pause innerhalb eines Satzes. Der Etikettierungsprozess umfasste Phrasen- und prosodische Informationen aus einem manuell getaggten Dungan-Text. Während dieser Phase überprüften und ergänzten Linguisten sporadisch ausgewählte Sätze. Durch iterative Korrekturen erreichten wir ein hohes Maß an Übereinstimmung mit Sprachexperten.

Trotz der Fähigkeit des BiLSTM, kontextabhängige Informationen zu lernen, werden seine unabhängigen Klassifizierungsentscheidungen durch starke Abhängigkeiten über das Ausgabelabel hinweg eingeschränkt. Um dies zu beheben, verwenden wir eine CRF-Schicht, die benachbarte Tags berücksichtigt, wie in Abbildung 4 dargestellt. Für einen normalisierten Eingabesatz **X** = {x1, x2, · · · , xn} mit n Wörtern und einer Tag-Sequenz des Satzes y = (y1, y2, . . . , yn) wird jedes Wort durch word2vec als d-dimensionaler Vektor dargestellt. Wir definieren seinen Vorhersagewert **s(X,** y) wie folgt:

$$\mathbf{s(X,}\ y) = \sum_{ich=1}^{N} P_{i,yi} + \sum_{ich=0}^{N} \text{Hallo}\ ,\ yi+1 \tag{1}$$

wobei **P** die Matrix der vom BLSTM-Netzwerk ausgegebenen Punktzahlen ist. Pi,yi entspricht der Punktzahl des yi- Tags des i-ten Wortes in einem Satz. **A** ist die Übergangspunktzahlmatrix der CRF-Schicht und Ayi ,yi+1 entspricht der Punktzahl von Tag yi zu Tag yi+1 .

Im Training maximieren wir die folgenden Log-Likelihood-Funktionen:

$$\log(p(y \mid X)) = \mathbf{s(X,}\ y) - \log \sum_{y \in YX} t^{s(X,y)} \tag{2}$$

wobei **YX** alle möglichen Tag-Sequenzen für einen Eingabetext X darstellt. ist wie

Bei der Dekodierung ist die optimale Sequenz y^{*} folgt gegeben: **s(X,** y) =

$$j^{*} = \underset{y \in YX}{\operatorname{argmax}} \tag{3}$$

2.1.5. Transformatorbasierte Konvertierung von Zeichen in Einheiten

Mandarin und Dunganisch verwenden dasselbe Pinyin-System zur Aussprachekennzeichnung. Folglich verläuft die Umwandlung von Zeichen in Einheiten im Dunganischen parallel zu der im Mandarin. Diese Studie stellt einen transformerbasierten Ansatz [43] zur Ableitung der Dungan-Einheit vor, wie in Abbildung 5 dargestellt, um die Genauigkeit der Umwandlung von Zeichen in Einheiten im Dunganischen zu verbessern. Der Encoder und Decoder werden durch Stapeln derselben wesentlichen Schichten mit N = 6 gebildet. Jede darunterliegende Schicht besteht aus zwei Unterschichten. Die erste Unterschicht ist die Multi-Head-Attention-Schicht. Der Decoder verfügt über eine Schicht versteckter Multi-Head-Attention (maskierte Multi-Head-Attention).

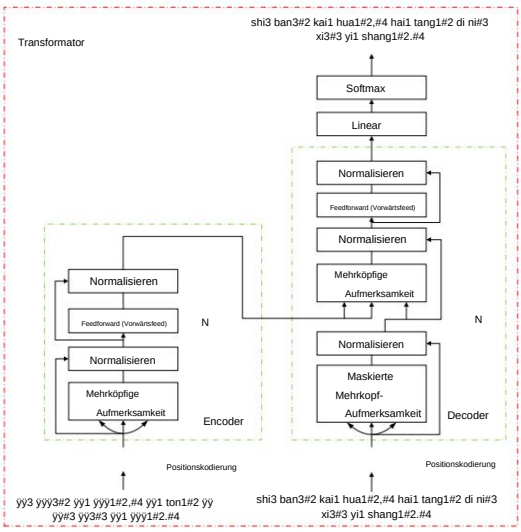

Abbildung 5. Das Framework der Transformer-basierten Konvertierung von Dungan-Zeichen in Einheiten. Die Eingabe ist ein Dungan-Satz mit prosodischen Informationen **(links)** und die entsprechende Pinyin-Sequenz **(rechts).** Die Ausgabe ist die Pinyin-Sequenz mit prosodischen Informationen.

2.2. Transferlernbasiertes Dungan Akustikmodell Wir

implementieren das Dungan Akustikmodell durch Feinabstimmung eines vorab trainierten Mandarin akustisches Modell, wie in Abbildung 6 dargestellt.

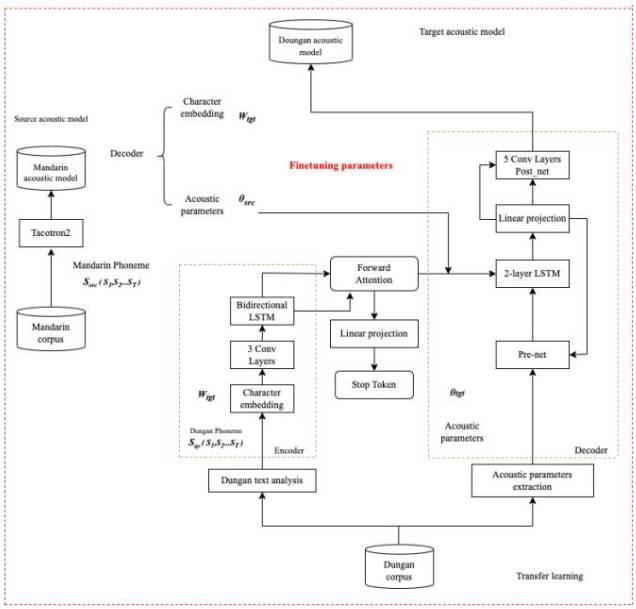

Abbildung 6. Verfahren zum Trainieren des akustischen Modells der Dungan-Sprache mit Transferlernen.

2.3. Vorab trainiertes, auf Tacotron2 basierendes Mandarin-Akustikmodell

Das Mandarin-Akustikmodell wird zunächst mithilfe eines umfangreichen Mandarin-Korpus trainiert. Unser proprietärer chinesischer Textanalysator extrahiert die Initialen, Endungen und zugehörigen prosodischen Bezeichnungen dieser Sätze. Die extrahierten akustischen Merkmale umfassen das Mel-Spektrogramm aus dem umfangreichen Mandarin-Korpus im Tacotron2-Framework.

Angesichts der ähnlichen Aussprache zwischen der Dungan-Sprache und Mandarin verwenden wir die Mapping-Transfer-Learning-Methode [44] , um durch Übertragung von Wissen aus dem Mandarin (Ausgangssprache) ein akustisches Modell für die Dungan-Sprache (Zielsprache) zu erhalten , das wie folgt formuliert werden kann :

$$f\ddot{y},W : XL\ \ddot{y}\ Y \tag{4}$$

wobei *ÿ* die Parameter des akustischen Modells sind, W lernbare Symboleinbettungen bezeichnet und Y den Raum des Mandarin darstellt. XL ist der Textraum für die Dungan-Sprache.

$$XL = \{st\} \quad \prod_{t=1}^{T} | \ddot{y}tst\ \ddot{y}\ L, T\ \ddot{y}\ N \tag{5}$$

Dabei ist L der Einheitensatz für die Dungan-Sprache, St die t-te Einheit der Dungan-Einheitenfolge und T die Länge der Einheitenfolge.

In den Encoder geben wir eine Dungan-Einheitensequenz ein, die durch Zeicheneinbettungen dargestellt wird. Diese wird durch einen Stapel aus drei Faltungsschichten geleitet, gefolgt von Batch- Normalisierung und ReLU-Aktivierungen. Anschließend wird die Ausgabe der letzten Faltungsschicht in eine bidirektionale LSTM-Schicht eingespeist, um die Dungan-Einheitsmerkmale zu generieren.

Beim Mapping-basierten Transferlernen werden Instanzen von ÿsrc und ÿtgt in einen neuen akustischen Parameterraum abgebildet. Dabei können wir Wsrc und ÿsrc, die vom Decoder aus dem Mandarin-Akustikmodell dekodiert wurden , direkt verwenden. ÿsrc und ÿtgt können Einbettungen als Eingabe verwenden und Sprache generieren. Da ssrc und stgt jedoch aus unterschiedlichen Symbolsätzen stammen, d. h. Lsrc = Ltgt, kann dasselbe Konzept nicht direkt auf Wsrc und Wtgt angewendet werden. Um dieses Problem zu lösen, werden Dunggan-Einheiten in Wtgt eingebettet, um das erneute Lernen während des Übertragungsprozesses zu erleichtern.

Wir übernehmen den Vorwärtsaufmerksamkeitsmechanismus, der kumulative Aufmerksamkeitsgewichte verwendet um den Kontextvektor zu berechnen.

Der Decoder ist ein autoregressives rekurrentes neuronales Netzwerk, das ein ÿtgt vorhersagt aus die Encoder-Eingangs-Dungan-Einheitssequenz ein Frame nach dem anderen. Wir können ÿsrc verwenden, das gelernt wurde aus dem Mandarin-Akustikmodell, um ÿtgt im neuen akustischen Parameterraum zu initialisieren. Die Ausgabe des ersten Zeitschritts wird zunächst durch ein Pre-Net verarbeitet, das aus zwei vollständig verbundene Schichten. Diese Ausgabe wird mit dem Vorwärtsaufmerksamkeitskontext kombiniert Vektor und durch ein Paar von LSTM-Schichten geleitet. Die Kombination der LSTM-Ausgaben und Aufmerksamkeitskontextvektoren durchlaufen drei verschiedene lineare Transformationen, um die Zielspektrogrammrahmen, Stopptoken und geschätzter Rest. Anschließend wird der vorhergesagte akustische Merkmale werden fünf Faltungsschichten unterzogen, wodurch ein Residuum zur Verbesserung erzeugt wird die Rekonstruktion des Dungan-Akustikmodells.

3. Ergebnisse

3.1. Auswertung der Transformer-Base-Dungan-Charakter-zu-Einheit-Konvertierung

Die Textanalyse im Frontend beeinflusst die Qualität der Sprachsynthese im Backend Ende, also haben wir den Dungan Textanalysator ausgewertet, bei dem die Konvertierung von Zeichen in Einheiten Modul ist der kritischste Faktor, der die Qualität der synthetisierten Sprache beeinflusst. Um zu beurteilen die Machbarkeit des transformerbasierten Dungan-Zeichen-zu-Einheit-Konvertierungsmoduls, verwendete einen Datensatz mit 10.783 Sätzen in der Dungan-Sprache, transkribiert mit Mandarin Pinyin. Die Dungan-Sprache und Mandarin Pinyin-Darstellungen des Datensatzes sind isomorph und umfassen Textattribute wie Ton und prosodische Grenzen, die zur Dungan-Sprache. In unserer Forschung haben wir 10 % der insgesamt 10.783 Sätze zugeordnet als Testdatensatz, weitere 10 % als Validierungsdatensatz und die restlichen 80 % wurden als Trainingssatz bezeichnet. Die mit dem Transformer verbundenen Hyperparameter sind detailliert in Tabelle 2. Wir verwendeten Präzision, Rückruf und F1-Maße als Bewertungsindizes, Die Ergebnisse des Evaluierungsprozesses bestätigten, dass die vorgeschlagenen Das Dungan Character-to-Unit-Modul eignet sich für die anschließende Auswertung der Sprachsynthese.

Tabelle 2. Die Hyperparameter des Transformer-basierten Zeichen-zu-Einheit-Konvertierungsmodells.

Parameter	Wert
Aufmerksamkeitsschichten Nx	6
Köpfe	8
Batchgröße	32
Versteckt	513
Ausfallen	0,1
Lernrate	0,0001

Tabelle 3. Die Ergebnisse der Transformer-basierten Konvertierung von Dungan-Zeichen in Einheiten.

Präzision	Abrufen	Formel 1
90.12	89,91	90.01

3.2. Evaluation von Dungan-Akustikmodellen auf Basis von Transferlernen

3.2.1. Korpus

Im Experiment nutzten wir Aufnahmen von neun weiblichen und einunddreißig männlichen Sprechern aus der Tsinghua Chinese 30-hour database [45] (insgesamt 13.389 Sätze) als Mandarin- Korpus. Für das Dungan-Korpus wählten wir Aufnahmen von fünf männlichen Sprechern aus (923 pro Person, insgesamt 4615 Sätze und 6 h). Das Dungan-Korpus umfasst alle anfänglichen und endgültige Aussprachen der Dungan-Sprache. Die durchschnittliche Satzlänge beträgt 18 Silben, mit einer durchschnittlichen Dauer von 10 s. Alle Aufnahmen wurden in ein einkanaliges 16 kHz Samplingfrequenz mit 16-Bit-Quantisierungsgenauigkeit.

3.2.2. Versuchsaufbau

Drei Arten von TTS-Frameworks, darunter Tacotron+Griffin-Lim, Tacotron2+WaveNet, und Tacotron2+WaveRNN, wurden in den Experimenten verglichen. Einige Hyperparameter von Die Rahmenbedingungen sind in Tabelle 4 aufgeführt.

Tabelle 4. Modell-Hyperparameter von Tacotron und Tacotron2.

Modell		**Tacotron**	**Tacotron2**	**Vorwärts-Achtung Tacotron2**
Vocoder		Griffin-Lim	WaveNet	WaveRNN
Encoder	Einbettung	Phomeme (256)	Phomeme (512)	Phomeme (512)
	Vornetz	FFN (256, 128)	-	FFN Phomeme (512, 256)
	Encoderkern	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)
	Aufmerksamkeit	Zusatzstoff (256)	Standortsensitiv (128)	Vorwärts (256)
	Achtung RNN	GRU (256)	LSTM (1024, 1024)	LSTM (256)
	Vornetz	FFN (256, 128)	FFN (256, 256)	FFN (256, 128)
Parameter		7,6 × 106	28,9 × 106	23,7 × 106

Alle drei Frameworks umfassen ein Frontend-Textanalysemodul, ein Akustikmodell Trainingsmodul und ein Vocoder. Das Textanalysemodul wandelt Dunganisch oder Chinesisch Sätze in eine durch Pinyin dargestellte Einheitssequenz, einschließlich Initialen, Endungen und deren Tönen und prosodische Grenzbezeichnungen. Im Trainingsmodul für akustische Modelle leiten wir den Logarithmus ab Magnitudenspektrogramm aus dem Sprachsignal mittels Hann-Fensterung mit einer 80 ms Framelänge, 12,5 ms Frameshift und eine 2048-Punkte-Fourier-Transformation.

Für das Tacotron+Griffin-Lim-Framework werden akustische Modelle mit einem Output trainiert Schichtreduktionsfaktor von r = 3 und der Adam-Optimierer mit abnehmender Lernrate. Der Die Lernrate beginnt bei 0,001 und wird anschließend auf 0,0005, 0,0003 und 0,0001 reduziert. nach 5, 20 und 50 Epochen. Eine einfache Verlustfunktion wird für die seq2seq-Decoder (Mel-Spektrogramm) und das Nachbearbeitungsnetzwerk (lineares Spektrogramm). Die Trainingsbatchgröße ist auf 32 eingestellt, wobei alle Sequenzen auf eine maximale Länge aufgefüllt werden durch Rekonstruktion der mit Nullen aufgefüllten Frames. Als Vocoder wird der Griffin-Lim-Algorithmus verwendet. zur Mel-Spektrum-zu-Sprache-Konvertierung.

Für das Tacotron2+WaveNet-basierte Framework trainieren wir die akustischen Modelle mit dem Standard-Maximum-Likelihood-Trainingsverfahren, bei dem die korrekte Ausgabe eingegeben wird anstelle der vorhergesagten Ausgabe auf der Decoderseite. Dies wurde mit einer Batchgröße abgeschlossen von 32. Der Adam-Optimierer wurde mit den folgenden Parametern verwendet: $\ddot{y}$ = 0,9, $\ddot{y}$ = 0,999, = 10ÿ6 . Die Lernrate wurde mit 10ÿ3 initialisiert und dann exponentiell auf 10ÿ5 reduziert. nach 50.000. Zusätzlich haben wir eine L2- Regularisierung mit einem Gewicht von 10ÿ6 angewendet . Für die Mel Für die Spektrum-zu-Sprache-Konvertierung wurde WaveNet als Vocoder eingesetzt.

In unserem Tacotron2+WaveRNN-basierten Transfer-Learning-Framework verwenden wir zunächst ein großes Mandarin-Korpus zum Vortrainieren eines Mandarin-Akustikmodells für nachfolgende Modelle Dieses vorab trainierte Modell wird dann verwendet, um das Dungan-Akustikmodell über den Transfer zu trainieren. Lernen aus dem Mandarin-Dungan-Korpus. Für die Vokodierung verwenden wir das WaveRNN für Mel-Spektrum-zu-Sprache-Konvertierung. Da Parametereinstellungen einen erheblichen Einfluss auf Modellgenauigkeit und Robustheit, wir haben diese Parameter durch iteratives Training optimiert und Updates.

Jedes TTS-Framework implementiert eine einsprachige Sprachsynthese für Mandarin oder Dunganisch und eine zweisprachige, die auf Transferlernen basiert. Wir haben mehrere Modelle in drei

TTS-Frameworks zur Beurteilung der Qualität und Klarheit der synthetisierten Sprache. In unserem Experiment 10% der Äußerungen wurden zufällig dem Testset zugeordnet, weitere 10% wurden für den Entwicklungssatz, und die verbleibenden Äußerungen bildeten den Trainingssatz.

Dunganisches einsprachiges Sprecher-abhängiges Modell

Wir trainierten das akustische Modell Dungan Monolingual Speaker-Dependent (DSD) mit Aufnahmen von fünf männlichen Sprechern, von denen jeder 923 Sätze beisteuerte, insgesamt 4615 Sätze und über 6 Stunden. Anschließend verglichen wir die Qualität und Klarheit der synthetisierten Sprache über drei Frameworks: DSD-Tacotron+Griffin-Lim, DSD Tacotron2+WaveNet, und DSD-Tacotron2+WaveRNN.

Modell für einsprachige Mandarin-Sprecher

Wir nutzten Aufnahmen von neun weiblichen und einunddreißig männlichen Sprechern (Tsinghua Chinesische 30-Stunden-Datenbank, bestehend aus 13.389 Sätzen) zum Trainieren des einsprachigen Mandarin Sprecherabhängiges (MSD) Akustikmodell. Wir verglichen die Qualität der synthetisierten Sprache und Klarheit über drei Frameworks: MSD-Tacotron+Griffin-Lim, MSD-Tacotron2+WaveNet, und MSD-Tacotron2+WaveRNN.

Mandarin und Dunganisch – zweisprachiges Sprecher-abhängiges Modell

Wir nutzten Aufnahmen von fünf männlichen Dungan-Sprechern (923 Sätze pro Person, 4615 Sätze, was 6 Stunden entspricht, werden als Trainingsdaten verwendet, um das Mandarin-Akustikmodell auf das Dungan-Akustikmodell zu übertragen und ein Dungan-Sprecher-abhängiges (MDSD) Akustikmodell und ein Mandarin Speaker-Dependent (MDSM) Akustikmodell. Wir Anschließend wurden die Qualität und Klarheit der synthetisierten Sprache in sechs Frameworks verglichen.

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

3.2.3. Objektive Bewertungen

Wir verwendeten die Mel-cepstrale Verzerrung (MCD) [46], Band A Periodicity Distortion (BAP) [47], mittlerer quadratischer Fehler (RMSE) [48] und Stimmloser/Stimmloser Fehler (V/UV) [47] um die verschiedenen Modelle objektiv zu bewerten. Die Ergebnisse für die Akustikmodelle DSD und MSD Modelle sind in Tabelle 5 bzw. Tabelle 6 dargestellt. Ebenso sind die MDSM- und MDSD-Die Ergebnisse werden in Tabelle 7 bzw. Tabelle 8 angezeigt.

Tabelle 5. Objektive Ergebnisse des DSD-Akustikmodells für Dungan.

Modell	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

Tabelle 6. Objektive Ergebnisse des MSD-Akustikmodells für Mandarin.

Modell	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

Tabelle 7. Objektive Ergebnisse des MDSD-Akustikmodells für Dungan.

Modell	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

Tabelle 8. Objektive Ergebnisse des MDSM-Akustikmodells für Mandarin.

Modell	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

Im Kontext der ressourcenarmen Dungan-Sprachsynthese beeinflusst die Qualität der Aufmerksamkeitsausrichtung zwischen Encoder und Decoder erheblich die Qualität der synthetisierten Sprache. Fehlausrichtungen zeigen sich vor allem in der Lesbarkeit, beim Überspringen und bei Wiederholungen. Daher verwenden wir die Diagonal Focus Rate (DFR) und die Word-Level Intelligibility Rate (IR) [49] zur Beurteilung der Lesbarkeit in ressourcenarmen Sprachen, wie in Tabelle 9 dargestellt. Der DFR stellt die Aufmerksamkeitskarte zwischen Encoder und Decoder dar und dient als architektonisches Metrik. Die IR misst den Prozentsatz der Testwörter, die richtig und deutlich ausgesprochen werden von Menschen, ein Standardmaß zur Beurteilung der Qualität der Sprachgenerierung mit geringen Ressourcen.

Tabelle 9. Lesbarkeit der synthetisierten Dungan-Sprache.

Modell	**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. Subjektive Bewertung

Zur subjektiven Bewertung wurden 30 Sätze zufällig aus dem Testset ausgewählt.
Wir haben drei Tests durchgeführt: Mean Opinion Score (MOS), Degradation Mean Opinion Score (DMOS) und AB-Präferenz zur Beurteilung der Qualität synthetisierter Sprache. Wir rekrutierten 20 Mandarin-Muttersprachler und 10 internationale Dungan-Muttersprachler (die verstanden Chinesisch) als Teilnehmer. Diese Teilnehmer erhielten vor der formellen Evaluierung eine Schulung. Die Mandarin-Teilnehmer beurteilten die Mandarin-Akustikmodelle von MSD und MDSM, während die Dungan-Teilnehmer die Dungan-Akustikmodelle von DSD und MDSD bewerteten. Beim MOS-Test bewerteten die Teilnehmer die Natürlichkeit der synthetisierten Sprache auf einer 5-Punkte-Skala. Skala. Die durchschnittlichen MOS-Werte für synthetisierte Dungan- und Mandarin-Sprache werden dargestellt in den Abbildungen 7 und 8.

Im DMOS-Test werden die synthetisierte Äußerung jedes Modells und die dazugehörige Originaläußerung Die Aufnahme bestand aus einem Paar Sprachdateien. Diese Paare wurden zufällig dem Probanden, wobei die synthetisierte Sprache der Originalsprache vorangestellt wurde. Die Teilnehmer wurden beauftragt mit dem sorgfältigen Vergleich der beiden Dateien und der Bewertung der Ähnlichkeit der synthetisierten Sprache mit dem Original auf einer 5-Punkte-Skala. Eine Punktzahl von 5 bedeutet, dass die synthetisierte Die Sprache war dem Original ähnlich, während ein Wert von 1 eine erhebliche Abweichung bedeutete. Die Abbildungen 9 und 10 zeigen die durchschnittlichen DMOS-Werte für synthetisierte Dungan- und Mandarin-Rede bzw.

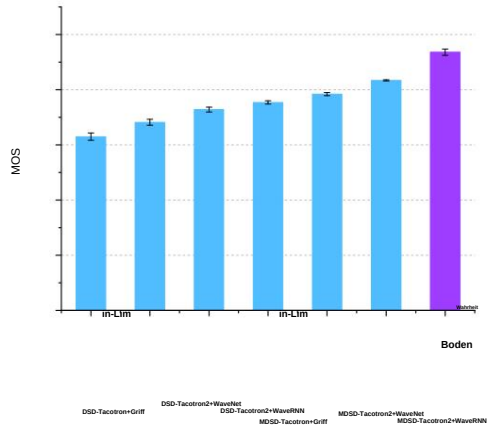

Abbildung 7. Die durchschnittlichen MOS-Werte der synthetisierten Dungan-Sprache unter 95ÿ%-Vertrauensintervallen.

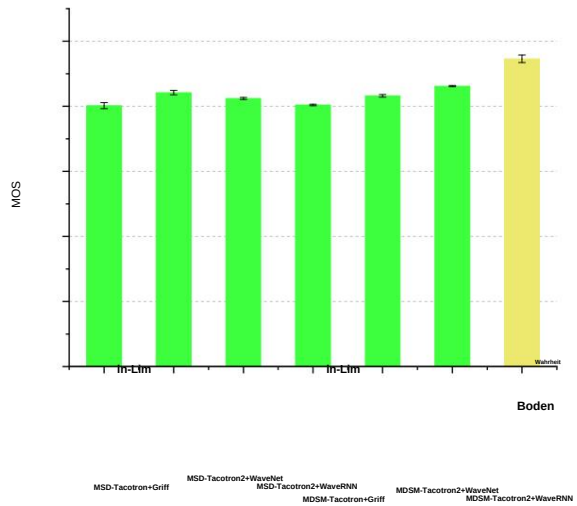

Abbildung 8. Die durchschnittlichen MOS-Werte der synthetisierten Mandarin-Sprache unter 95ÿ%-Konfidenzintervallen.

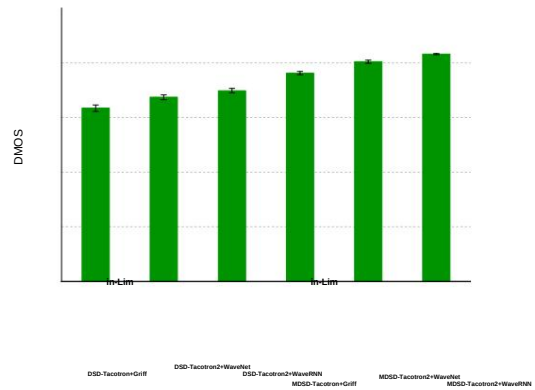

Abbildung 9. Die durchschnittlichen DMOS-Werte der synthetisierten Dungan-Sprache unter 95ÿ%-Vertrauensintervallen.

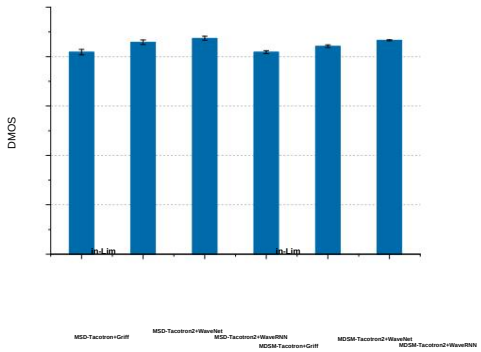

Abbildung 10. Die durchschnittlichen DMOS-Werte der synthetisierten Mandarin-Sprache unter 95ÿ%-Konfidenzintervallen.

Im AB-Präferenztest bestand jedes Paar aus zwei identischen Sätzen. Die synthetisierten Äußerungen wurden in zufälliger Reihenfolge abgespielt. Die Teilnehmer wurden angewiesen, zuzuhören und bewerten Sie, welche Äußerung die bessere Qualität hatte oder geben Sie „neutral“ an, wenn keine Präferenz vorliegt wurde erkannt. Die synthetisierten Dungan- und Mandarin-Sprachpräferenzergebnisse sind dargestellt in Tabelle 10 bzw. Tabelle 11 .

Tabelle 10. Subjektiver AB-Präferenzwert (%) von Dungan mit *ÿ* < 0,01.

	DSD-Tacotron+ Griffin-Lim	DSD-Tacotron2+ WaveNet	DSD-Tacotron2 +WellenRNN	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

Tabelle 11. Subjektiver AB-Präferenzwert (%) für Mandarin mit *ÿ* < 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. Diskussion

In objektiven Bewertungen, obwohl das Tacotron+Griffin-Lim-basierte TTS-Framework ordnet linguistische Merkmale akustischen Merkmalen Bild für Bild durch das einsprachige Dungan -Korpus zu. Die Qualität und Lesbarkeit der synthetisierten Dungan-Sprache muss verbessert werden. Die nach vorn gerichtete Aufmerksamkeit und ein fein abgestimmtes akustisches Modell können jedoch die Lesbarkeit verbessern und die Trainingszeit zu verkürzen. Folglich ist das auf Transferlernen basierende Tacotron2+WaveRNN Das Akustikmodell des Frameworks übertrifft andere. Die objektiven Ergebnisse des MDSD-Akustikmodells übertreffen die des DSD-Akustikmodells. Dies liegt daran, dass Dungan eine Variation ist des nordwestlichen chinesischen Dialekts, der viele interne Ähnlichkeiten aufweist. Angesichts der Ausspracheähnlichkeiten zwischen Mandarin und Dunganisch repräsentiert das gleiche Symbol ihre genaue Aussprachen. Daher kommen wir zu dem Schluss, dass das Hinzufügen eines Mandarin-Korpus und die Verwendung Transferlernen kann die Qualität und Lesbarkeit der synthetisierten Dungan-Sprache verbessern.

Alle subjektiven Bewertungen stimmen in verschiedenen Aspekten mit objektiven Einschätzungen überein. Das auf Transferlernen basierende Tacotron2+waveRNN-Framework bietet eine hervorragende Sprachqualität,

insbesondere in Bezug auf die Natürlichkeit und Lesbarkeit der synthetisierten Sprache. Mit der Ergänzung des Mandarin-Korpus übertreffen die Qualität und Lesbarkeit der synthetisierten Dungan-Sprache unter Verwendung der auf Transferlernen basierenden TTS-Frameworks die der monolingualen, mit Korpus trainierten TTS-Frameworks. Dies wird durch den AB-Präferenztest weiter bestätigt, der bestätigt, dass unsere vorgeschlagenen TTS-Frameworks im Vergleich zur durch das monolinguale Akustikmodell synthetisierten Sprache eine verbesserte Qualität und Lesbarkeit bieten.

5. Schlussfolgerungen

Diese Studie erweitert unsere bisherige Forschung durch die Implementierung einer auf Transferlernen basierenden Man- darin-Sprachsynthese und einer ressourcenarmen Dungan-Sprachsynthese im Rahmen von Tacotron2+WaveRNN. Wir haben außerdem einen umfassenden Dungan-Textanalysator entwickelt. Objektive und subjektive Experimente haben gezeigt, dass die auf Transferlernen basierende Dungan-Sprachsynthese im Rahmen von Tacotron2+WaveRNN alternative Methoden und das einsprachige Dungan-Sprachsynthese-Framework übertraf. Darüber hinaus beeinträchtigte das Transferlernen weder die Sprachqualität noch die Lesbarkeit der synthetisierten ressourcenarmen Dungan-Sprache. Daher birgt unser Ansatz erhebliches Potenzial für die Entwicklung von Sprachsynthesesystemen für ressourcenarme Minderheitensprachen.

Im Bereich TTS wurden auf der Grundlage tiefer neuronaler Netzwerke zahlreiche Durchbrüche erzielt. Uns ist aufgefallen, dass in letzter Zeit einige neue Methoden zur Sprachsynthese [50–52] vorgeschlagen wurden. Motiviert durch die jüngsten Fortschritte bei autoregressiven (AR) Modellen, die reine Decoder-Architekturen zur Textgenerierung verwenden, wenden mehrere Studien, wie VALL-E [53] und BASE TTS [54], ähnliche Architekturen auf TTS-Aufgaben an. Diese Studien demonstrieren die bemerkenswerte Fähigkeit reiner Decoder-Architekturen, natürlich klingende Sprache zu erzeugen.
Diese Studien demonstrieren die bemerkenswerte Fähigkeit von reinen Decoder-Architekturen, natürlich klingende Sprache zu produzieren. Zukünftige Forschung wird sich darauf konzentrieren, diese neuen Methoden zu nutzen, um die Qualität der Sprachsynthese der Dungan-Sprache zu verbessern, die Größe des Dungan-Korpus zu reduzieren und Sprachsynthese für Dungan-Sprachen mit einem größeren Korpus zu erreichen. Darüber hinaus wird Multitask-Lernen erforscht, um sprecherunabhängige Szenarien zu realisieren und die Emotionalität der synthetisierten Dungan-Sprache zu verbessern.

Beiträge der Autoren: Konzeptualisierung, ML und HY; formale Analyse, HY und RJ; Datenkuratierung , ML und RJ; Schreiben – Vorbereitung des Originalentwurfs, ML und RJ; Schreiben – Überprüfung und Bearbeitung, HY und ML; Überwachung, HY; Mittelbeschaffung, HY. Alle Autoren haben die veröffentlichte Version des Manuskripts gelesen und stimmen ihr zu .

Finanzierung: Die Forschung wird durch den Forschungsfonds der National Natural Science Foundation of China (Zuschuss-Nr. 62067008) unterstützt.

Erklärung des Institutional Review Board: Gilt nicht für Studien, an denen keine Menschen oder Tiere beteiligt sind.

Einverständniserklärung: Nicht zutreffend.

Datenverfügbarkeitserklärung: Wir haben im Manuskript zwei Trainingsdatensätze verwendet. Einer ist ein öffentlich verfügbarer Mandarin-Datensatz (THCHS-30) und der andere ist ein Donggan-Datensatz, der Sprache und Text enthält. Ersterer ist öffentlich und kann unter http://www.openslr.org/18/ abgerufen werden. (abgerufen am 16. Juni 2024). Letzterer ist ein selbst erstellter Datensatz und nicht öffentlich zugänglich. Die Daten werden jedoch auf Anfrage zur Verfügung gestellt.

Interessenkonflikte: Die Autoren erklären, dass kein Interessenkonflikt besteht. Die Geldgeber hatten keinen Einfluss auf die Gestaltung der Studie, auf die Erhebung, Analyse oder Interpretation der Daten, auf das Schreiben des Manuskripts oder auf die Entscheidung, die Ergebnisse zu veröffentlichen.

Verweise

1. Tu, T.; Chen, YJ; Chieh Yeh, C.; Yi Lee, H. End-to-End-Text-to-Speech für ressourcenarme Sprachen durch sprachübergreifende Übertragung Lernen. arXiv **2019,** arXiv:1904.06508.
2. Liu, R.; Sisman, B.; Bao, F.; Yang, J.; Gao, G.; Li, H. Nutzung morphologischer und phonologischer Merkmale zur Verbesserung der prosodischen Phrasierung für die mongolische Sprachsynthese. 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. Textinduktive, Graphon-basierte Sprachadaption für ressourcenarme Sprachsynthese. 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, TY LRSpeech: Sprachsynthese und -erkennung mit extrem geringem Ressourcenaufwand. In Proceedings der 26. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'20, New York, NY, USA, 6.–10. Juli 2020; S. 2802–2812. [Querverweis]
5. He, M.; Yang, J.; He, L.; Soong, FK Mehrsprachige Byte2Speech-Modelle für skalierbare ressourcenarme Sprachsynthese. arXiv **2021,** arXiv:2103.03541.
6. Oliveira, FS; Casanova, E.; Junior, AC; Soares, AS; Galvão Filho, AR CML-TTS: Ein mehrsprachiger Datensatz für die Sprachsynthese in ressourcenarmen Sprachen. In Text, Sprache und Dialog; Ekštein, K., Pártl, F., Konopík, M., Hrsg.; Springer: Cham, Schweiz, 2023; S. 188–199.
7. Zhu, Y. Donggan-Sprache: Eine besondere Variante der Shaanxi- und Gansu-Dialekte. Asian Lang. Cult. **2013,** 4, 51–60. 8.
Jiang, Y. Die Donggan-Sprache und ihre Beziehung zu den Shaanxi- und Gansu-Dialekten. J. Chin. Linguist. **2014,** 42, 229–258.
9. Chen, L.; Yang, H.; Wang, H. Forschung zur Dungan-Sprachsynthese basierend auf Deep Neural Network. In Proceedings des 11. Internationalen Symposiums zur Verarbeitung chinesischer gesprochener Sprache (ISCSLP) 2018, Taipeh, Taiwan, 26.–29. November 2018; S. 46–50. [CrossRef]
10. Jiang, R.; Chen, C.; Shan, X.; Yang, H. Sprachverbesserung zur Realisierung der Sprachsynthese ressourcenarmer Dungan-Sprachen. In Proceedings der 24. Konferenz des Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) 2021, Singapur, 18.–20. November 2021; S. 193–198. [CrossRef]
11. Hunt, AJ; Black, AW Einheitenauswahl in einem konkatenativen Sprachsynthesesystem unter Verwendung einer großen Sprachdatenbank. In Proceedings der 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9. Mai 1996; Band 1, Seiten 373–376.
12. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Sprachsynthese basierend auf versteckten Markov-Modellen. Proc. IEEE **2013,** 101, 1234–1252. [CrossRef]
13. Ling, ZH; Deng, L.; Yu, D. Modellierung von Spektralhüllkurven mit eingeschränkten Boltzmann-Maschinen und tiefen Belief-Netzwerken für die statistische parametrische Sprachsynthese. IEEE Trans. Audio Speech Lang. Process. **2013,** 21, 2129–2139. [CrossRef]
14. Zen, H.; Senior, A.; Schuster, M. Statistische parametrische Sprachsynthese mit tiefen neuronalen Netzwerken. In Proceedings der 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Kanada, 26.–31. Mai 2013; S. 7962–7966. [CrossRef]
15. Wang, P.; Qian, Y.; Soong, FK; He, L.; Zhao, H. Word Embedding für rekurrente neuronale Netze basierende TTS-Synthese. In Proceedings der 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australien, 19.–24. April 2015; S. 4879–4883. [CrossRef]
16. Yu, Q.; Liu, P.; Wu, Z.; Ang, SK; Meng, H.; Cai, L. Lernen sprachübergreifender Informationen mit mehrsprachigem BLSTM für die Sprachsynthese ressourcenarmer Sprachen. In Proceedings der 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20.–25. März 2016; S. 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-Synthese mit menschlicher Qualität. IEEE Trans. Pattern Anal. Mach. Intell. **2024,** 46, 4234–4245. [CrossRef]
18. Wang, Y.; Skerry-Ryan, RJ; Stanton, D.; Wu, Y.; Weiss, RJ; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Auf dem Weg zur End-to-End-Sprachsynthese. In Proceedings der 18. Jahreskonferenz der International Speech Communication Association, Interspeech 2017, Stockholm, Schweden, 20.–24. August 2017.
19. Shen, J.; Pang, R.; Weiss, RJ; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natürliche TTS-Synthese durch Konditionierung von Wavenet auf MEL-Spektrogrammvorhersagen. In Proceedings der 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Kanada, 15.–20. April 2018; S. 4779–4783. [CrossRef]
20. Griffin, D.; Lim, J. Signalschätzung aus modifizierter Kurzzeit-Fourier-Transformation. IEEE Trans. Acoust. Sprachsignalprozess. **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: Ein generatives Modell für Rohaudio. 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. Effiziente neuronale Audiosynthese. arXiv **2018,** arXiv:1802.08435.
23. Byambadorj, Z.; Nishimura, R.; Ayush, A.; Ohta, K.; Kitaoka, N. Text-to-Speech-System für ressourcenarme Sprachen unter Verwendung von sprachübergreifendem Transferlernen und Datenerweiterung. EURASIP J. Audio Speech Music. Prozess. **2021,** 2021, 42. [CrossRef]
24. Joshi, R.; Garera, N. Schnelle Sprecheranpassung in ressourcenarmen Text-to-Speech-Systemen unter Verwendung synthetischer Daten und Transferlernen . In Proceedings der 37. Pacific Asia Conference on Language, Information and Computation, Hongkong, China, 2.–4. Dezember 2023; Huang, CR, Harada, Y., Kim, JB, Chen, S., Hsu, YY, Chersoni, EAP, Zeng, WH, Peng, B., Li, Y., et al., Hrsg.; ACL: Hongkong, China, 2023; S. 267–273.
25. Do, P.; Coler, M.; Dijkstra, J.; Klabbers, E. Strategien des Transferlernens für ressourcenarme Sprachsynthese: Phone Mapping, Feature-Input und Auswahl der Ausgangssprache. In Proceedings des 12. ISCA Speech Synthesis Workshop (SSW2023), Grenoble, Frankreich, 26.–28. August 2023; S. 21–26. [CrossRef]

26. Azizah, K.; Jatmiko, W. Transferlernen, Stilkontrolle und Sprecherrekonstruktionsverlust für Zero-Shot-mehrsprachige Mehrsprecher Text-to-Speech in ressourcenarmen Sprachen. IEEE Access **2022,** 10, 5895–5911. [CrossRef]
27. Cai, Z.; Yang, Y.; Li, M. Sprachübergreifende Mehrsprecher-Sprachsynthese mit begrenzten zweisprachigen Trainingsdaten. Comput. Speech Lang. **2023,** 77, 101427. [CrossRef]
28. Yang, H.; Oura, K.; Wang, H.; Gan, Z.; Tokuda, K. Verwendung von adaptivem Sprechertraining zur Realisierung von Mandarin-Tibetisch-übergreifendem Sprachsynthese. Multimed. Tools Appl. **2015,** 74, 9927–9942. [CrossRef]
29. Wang, L.; Yang, H. Tibetische Wortsegmentierungsmethode basierend auf dem bilstm_crf-Modell. In Proceedings der IEEE 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesien, 15.–17. November 2018; S. 297–302.
30. Zhang, W.; Yang, H.; Bu, X.; Wang, L. Deep Learning für die sprachenübergreifende Mandarin-Tibetan-Sprachsynthese. IEEE Access **2019,** 7, 167884–167894. [CrossRef]
31. Zhang, W.; Yang, H. Verbesserung der Sequenz-zu-Sequenz-Sprachsynthese im Tibetischen mit prosodischen Informationen. ACM Trans. Asian Low-Resour. Lang. Inf. Process. **2023,** 22, 6012. [CrossRef]
32. Zhang, W.; Yang, H. Meta-Learning für die sprachübergreifende Mandarin-Tibetan-Sprachsynthese. Appl. Sci. **2022,** 12, 2185. [CrossRef]
33. Hai, F. Eine Pilotstudie zu Lehnwörtern in der zentralasiatischen Dungan-Sprache. Xinjiang Univ. J. **2000,** 28, 58–63.
34. Lin, T. Merkmale, Situation und Entwicklungstrends der Tung'gan-Sprache in Zentralasien. Contemp. Linguist. **2016,** 18, 234–243.
35. Gladney, DC Relationale Alterität: Konstruktion dunganischer (hui), uigurischer und kasachischer Identitäten in China, Zentralasien und der Türkei. Hist. Anthropol. **1996,** 9, 445–477. [CrossRef]
36. Miao, DX Zweisprachiges Unterrichtsmodell des Donggan-Volkes. J. Res. Educ. Ethn. Minor. **2008,** 19, 111–114.
37. Jia, Y.; Huang, D.; Liu, W.; Dong, Y.; Yu, S.; Wang, H. Textnormalisierung in einem Mandarin-Text-to-Speech-System. In Proceedings der 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31. März–4. April 2008; S. 4693–4696. [CrossRef]
38. Wanmezhaxi, N. Forschung zu mehreren Schlüsselfragen der tibetischen Wortsegmentierung. J. Chin. Inf. Process. **2014,** 28, 132–139.
39. Zavyalova, O. Dungan Language. 2015. Online verfügbar: https://www.academia.edu/42869092/Dungan_Language (Zugriff am 16. Juni 2024).
40. Lin, T. Donggan-Schrift – Ein erfolgreicher Versuch der chinesischen alphabetischen Schrift. J. Second. Northwest Univ. Natl. **2005,** 2005, 31–36.
41. Yang, WJ; Zhang, R. Ethnische Identität im Kontext mehrerer Nationen – Eine Fallstudie zu „Dunggan“ und der Hui-Nationalität. J. South-Cent. Univ. Natl. **2009,** 29, 31–36.
42. Zheng, Y.; Tao, J.; Wen, Z.; Li, Y. BLSTM-CRF-basierte End-to-End-Prosodische-Grenzwert-Vorhersage mit kontextsensitiven Einbettungen in einem Text-to-Speech-Frontend. Proc. Interspeech **2018,** 9, 47–51. [CrossRef]
43. Hlaing, AM; Pa, WP Sequenz-zu-Sequenz-Modelle für die Umwandlung von Graphemen in Phoneme im großen myanmarischen Aussprachewörterbuch . In Proceedings der 22. Konferenz des Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Cebu, Philippinen, 25.–27. Oktober 2019; S. 1–5. [CrossRef]
44. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. Eine Umfrage zum Deep Transfer Learning. In Künstliche neuronale Netzwerke und maschinelles Lernen – ICANN 2018; K ŷurková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I., Hrsg.; Springer: Cham, Schweiz, 2018; S. 270–279.
45. Wang, D.; Zhang, X. THCHS-30: Ein freies chinesisches Sprachkorpus. arXiv **2015,** arXiv:1512.01882.
46. Kubichek, R. Mel-cepstrale Distanzmessung zur objektiven Beurteilung der Sprachqualität. In Proceedings der IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Victoria, BC, Kanada, 19.–21. Mai 1993; Band 1, Seiten 125–128. [Querverweis]
47. Dhiman, JK; Seelamantula, CS Eine spektrotemporale Technik zur Schätzung von Aperiodizität und stimmhaften/stimmlosen Entscheidungsgrenzen von Sprachsignalen. In Proceedings der 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019), Brighton, Großbritannien, 12.–17. Mai 2019; S. 6510–6514. [CrossRef]
48. Castelazo, I.; Mitani, Y. Zur Verwendung des mittleren quadrierten Fehlers als Leistungsindex. Accredit. Qual. Assur. **2012,** 17, 95–97. [Querverweis]
49. Ren, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, TY Fast unbeaufsichtigte Text-to-Speech- und automatische Spracherkennung. arXiv **2020,** arXiv:1905.06791.
50. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, TY FastSpeech 2: Schnelle und qualitativ hochwertige End-to-End-Text-to-Speech-Umwandlung. arXiv **2022,** arXiv:2006.04558.
51. Chen, J.; Song, X.; Peng, Z.; Zhang, B.; Pan, F.; Wu, Z. LightGrad: Leichtgewichtiges Diffusions-Probabilistikmodell für Text-to-Speech. In Proceedings der 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2023), Rhodos , Griechenland, 4.–10. Juni 2023; S. 1–5. [CrossRef]
52. Guo, Y.; Du, C.; Ma, Z.; Chen, X.; Yu, K. VoiceFlow: Effiziente Text-to-Speech-Technologie mit Rectified Flow Matching. In Proceedings der 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2024), Seoul, Republik Korea, 14.–19. April 2024; S. 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. Neuronale Codec-Sprachmodelle sind Zero-Shot-Text-to-Speech-Synthesizer. arXiv **2023,** arXiv:2301.02111.
54. ÿajszczak, 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: Lehren aus dem Aufbau eines Text-to-Speech-Modells mit einer Milliarde Parametern auf Grundlage von 100.000 Stunden Daten. arXiv **2024,** arXiv:2402.08093.