sciences appliquées

MDPI

Article

# Utiliser l'apprentissage par transfert pour réaliser un Dungan à faibles ressources Synthèse vocale du langage

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

1 Collège d'ingénierie électronique et de l'information, Université Tongji, Shanghai 201804, Chine ; liuxh709@163.com

2 École de technologie éducative, Université normale du Nord-Ouest, Lanzhou 730070, Chine ; jiangh940618@163.com

3 Laboratoire clé de numérisation de l'éducation de la province du Gansu, Lanzhou 730070, Chine

* Correspondance : yanghw@nwnu.edu.cn

† Ces auteurs ont participé à ce travail à part égale.

Résumé : Cet article présente une méthode basée sur l'apprentissage par transfert pour améliorer la qualité de la parole synthétisée de la langue Dungan à faibles ressources. Cette amélioration est réalisée en ajustant avec précision un modèle acoustique en mandarin pré-entraîné à un modèle acoustique en langue Dungan en utilisant un corpus Dun-gan limité dans le cadre Tacotron2+WaveRNN. Notre méthode commence par le développement d'un analyseur de texte Dungan basé sur un transformateur, capable de générer des séquences d'unités avec des informations prosodiques intégrées à partir de phrases Dungan. Ces séquences unitaires, ainsi que les caractéristiques vocales, fournissent des paires <séquence unitaire avec étiquettes prosodiques, spectrogrammes Mel> comme entrée de Tacotron2 pour entraîner le modèle acoustique. Parallèlement, nous avons pré-entraîné un modèle acoustique mandarin basé sur Tacotron2 en utilisant un corpus mandarin à grande échelle. Le modèle est ensuite affiné avec un corpus vocal Dungan à petite échelle pour dériver un modèle acoustique Dungan qui apprend de manière autonome l'alignement et la cartographie des unités avec les spectrogrammes. Les spectrogrammes résultants sont convertis en formes d'onde via le vocodeur WaveRNN, facilitant la synthèse de discours mandarin ou Dungan de haute qualité. Des expériences subjectives et objectives suggèrent que la synthèse vocale Dungan basée sur l'apprentissage par transfert proposée obtient des scores supérieurs par rapport aux modèles formés uniquement avec le corpus Dungan et d'autres méthodes. Par conséquent, notre méthode propose une stratégie pour réaliser la synthèse vocale pour les langues à faibles ressources en ajoutant des informations prosodiques et en exploitant un corpus linguistique similaire à hautes ressources grâce à l'apprentissage par transfert.

Mots-clés : synthèse vocale en langue Dungan ; analyse de texte ; apprentissage par transfert ; langue à faibles ressources ; tacotron2



Citation : Liu, M. ; Jiang, R. ; Yang, H. Utiliser l'apprentissage par transfert pour réaliser une synthèse vocale en langage Dungan à faibles ressources. Appl. Sci. 2024, 14, 6336. https://doi.org/10.3390/app14146336

Rédactrices académiques : Gloria Corpas Pasteur et Tharindu Ranasinghe

Reçu : 17 juin 2024
Révisé : 17 juillet 2024
Accepté : 18 juillet 2024
Publié : 20 juillet 2024



## 1. Introduction

La synthèse vocale (conversion texte-parole (TTS)) est largement utilisée dans les maisons intelligentes, les systèmes de navigation et les applications de livres audio. Il existe dans le monde environ 6 000 langues , la plupart considérées comme à faibles ressources. Bien que des progrès significatifs aient été réalisés dans la synthèse vocale pour les principales langues comme le mandarin, l'anglais et le français, la qualité vocale de la TTS pour les langues à faibles ressources, comme le tibétain et le doungan, reste sous-optimale. Ces dernières années, il y a eu une recrudescence des recherches axées sur la synthèse vocale des langues à faibles ressources, comme en témoignent de nombreuses études [1–6]. Cependant, les recherches sur la synthèse vocale du langage Dungan doivent encore être complétées. La langue Dungan, qui est une variante des dialectes Shanxi-Gansu au sein du dialecte chinois parlé en Asie centrale, est classée comme une langue à faibles ressources en raison de son utilisation limitée, de la diminution du nombre de locuteurs et de la rareté du matériel linguistique [7, 8]. Étant donné que le russe est devenu la langue officielle de l'Asie centrale, la création d'un corpus vocal complet doté de connaissances linguistiques pour une synthèse vocale Dungan de haute qualité présente un défi de taille. Même si n

Synthèse vocale Dungan basée sur DNN [9,10], la qualité de la parole synthétisée n'était pas élevée en raison du corpus de formation limité.

Les technologies de synthèse vocale englobent la synthèse vocale concaténative basée sur la sélection d'unités [11], la synthèse vocale paramétrique statistique (SPSS) basée sur un modèle de Markov caché (HMM ) [12] et la synthèse vocale basée sur l'apprentissage profond [13,14]. Alors que l'apprentissage profond a considérablement avancé la technologie de synthèse vocale, des méthodes telles que la mémoire à long terme (LSTM) et le LSTM bidirectionnel [15,16] ont résolu les limitations de l'information temporelle. De plus, des modèles de synthèse vocale de bout en bout [17] comme Tacotron [18] et Tacotron2 [19] ont démontré la capacité de mapper le texte directement à la parole. Lorsqu'ils sont entraînés avec des paires texte-parole à grande échelle, ces modèles produisent une parole synthétisée à l'aide de vocodeurs de haute qualité tels que l'algorithme Griffin-Lim [20], WaveNet [21] et WaveRNN [22].
De tels systèmes nécessitent cependant des corpus de formation conséquents. Pour les langues à faibles ressources, le manque de corpus de formation rend difficile pour les modèles de bout en bout d'apprendre la structure prosodique des phrases, ce qui entraîne un manque de changements prosodiques dans la parole synthétisée, ce qui affecte son caractère naturel, posant des défis pour la synthèse vocale. des langues à faibles ressource

L'apprentissage par transfert entre langues [23-25] a été utilisé pour atténuer le problème de l'insuffisance des corpus de formation pour la synthèse vocale dans les langues à faibles ressources. Cette technique implique la formation d'un modèle de langage en utilisant une combinaison d'un grand corpus d'un langage à ressources élevées et d'un corpus plus petit d'un langage à faibles ressources, suivi de l'adaptation de ce modèle au langage à faibles ressources. L'apprentissage par transfert en synthèse vocale s'est avéré être une stratégie efficace pour produire de la parole dans des langues à faibles ressources en exploitant les capacités d'un modèle acoustique de langue à hautes ressources [26,27].

Dans nos recherches antérieures sur la synthèse vocale tibétaine [28-32], nous avons déterminé que l'intégration d'informations prosodiques via des techniques basées sur l'apprentissage par transfert améliore la qualité de la parole synthétisée pour les langues à faibles ressources telles que le tibétain. S'appuyant sur cette idée, la présente étude met en œuvre une approche séquence à séquence (seq2seq) pour la synthèse vocale en langage Dungan, tirant parti de l'apprentissage par transfert et des informations prosodiques dans le cadre Tacotron2+WaveRNN. Cette méthode implique l'utilisation d'un analyseur de texte Dungan pour extraire les étiquettes prosodiques des phrases Dungan pour l'intégration de modèles, en utilisant un modèle acoustique en mandarin basé sur Tacotron2 et en affinant le modèle acoustique de la langue Dungan avec un corpus vocal Dungan limité. Les principales contributions sont décrites ci-dessous :

- Front-end : nous avons implémenté un analyseur de texte complet pour le langage Dungan, comprenant des modules de normalisation de texte, de segmentation de mots, de prédiction des limites prosodiques et de génération d'unités basées sur la technologie des transformateurs. Cet analyseur peut produire des initiales et des finales sous forme d'unités de synthèse vocale avec des étiquettes prosodiques à partir de phrases Dungan. •
- Back-end : nous avons réalisé la synthèse vocale seq2seq Dungan en adaptant un modèle acoustique mandarin pré-entraîné dans le cadre Tacotron2+WaveRNN.
  Ceci a été accompli en remplaçant l'attention sensible à la localisation de Tacotron2 par une attention vers l'avant, améliorant ainsi la vitesse et la stabilité de la convergence.

Le reste de l'article est organisé comme suit. Nous présentons d'abord notre cadre de synthèse vocale Dungan basé sur l'apprentissage par transfert sous Tacotron2+WaveRNN dans la section 2. La configuration expérimentale et les résultats sont présentés dans la section 3, tandis que les résultats sont discutés dans la section 4. Enfin, une brève conclusion et un aperçu des travaux futurs. sont fournis dans la section 5

## 2. Modèles et méthodes

Le cadre proposé pour la synthèse vocale Dungan à faibles ressources basée sur l'apprentissage par transfert , illustré à la figure 1, comprend un module d'extraction de caractéristiques, un modèle acoustique mandarin pré-entraîné , un module de formation de modèle acoustique Dungan basé sur l'apprentissage par transfert et un synthétiseur vocal basé sur un vocodeur WaveRNN.

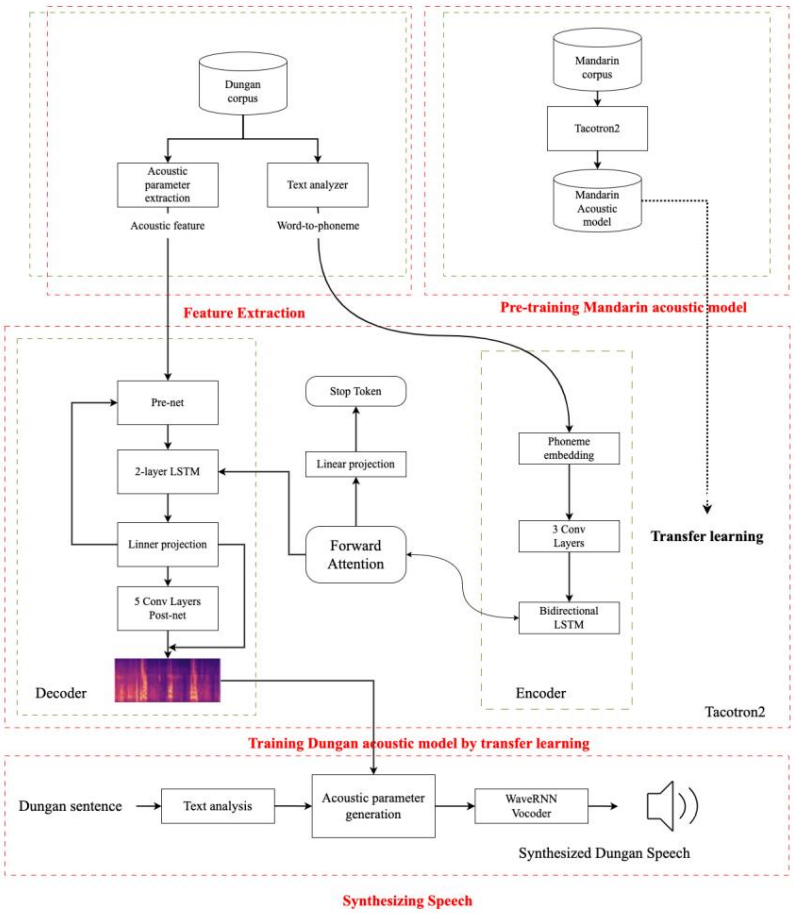


Figure 1. Le cadre de la synthèse vocale Dungan basée sur Tacotron2+WaveRNN.

Le module d'extraction de caractéristiques extrait les caractéristiques acoustiques telles que le spectrogramme de Mel à partir de signaux vocaux et la séquence d'unités de synthèse vocale à partir de phrases. Nous avons développé un analyseur de texte complet en langue Dungan pour extraire des unités de synthèse vocale avec des caractéristiques prosodiques afin de mapper les phrases Dungan sur des séquences d'unités. Étant donné que les langues mandarin et dungan utilisent les initiales et les finales comme unités principales de synthèse vocale, la séquence d'unités résultante incorpore ces éléments et des informations prosodiques pertinentes, y compris les tons des syllabes et les étiquettes de limites prosodiques au niveau des phrases.

Étant donné que Tacotron2 est l'un des cadres de synthèse vocale codeur-décodeur les plus populaires et que le vocodeur WaveRNN peut générer une parole naturelle, nous utilisons Tacotron2 pour entraîner des modèles acoustiques et WAVRNN pour convertir le spectrogramme en forme d'onde pour la langue Dungan et le mandarin. Le modèle acoustique mandarin est pré-entraîné avec un corpus mandarin à grande échelle, tandis que le modèle linguistique Dungan est transféré du modèle acoustique mandarin avec un corpus Dungan à petite échelle.

Au stade de la synthèse vocale, le vocodeur WaveRNN génère un discours en dungan ou en mandarin à partir de l'entrée de phrases en dungan ou en chinois. L'analyseur de texte génère d'abord les étiquettes dépendant du contexte à partir de la phrase d'entrée. Ensuite, les séquences d'unités de synthèse vocale (initiales et finales avec leurs informations prosodiques) sont introduites dans le modèle acoustique mandarin ou Dungan pour générer le spectrogramme Mel. Le vocodeur WaveRNN est enfin utilisé pour générer les formes d'onde vocales à partir du spectrogramme Mel. Nous utilisons un analyseur de texte chinois développé en interne pour l'analyse de texte chinois.

### 2.1. Analyseur de texte de la langue Dungan

Contrairement aux techniques de synthèse vocale seq2seq les plus répandues, conçues pour les principales langues et qui utilisent uniquement la paire <séquence phonème, parole> pour entraîner des modèles acoustiques, notre approche utilise une séquence unitaire incorporant des étiquettes prosodiques telles que le

le ton de chaque syllabe et la limite prosodique d'une phrase, servant de « séquence phonétique ». Par conséquent, il devient essentiel de concevoir un analyseur de texte complet capable d'extraire les séquences unitaires d'une phrase et leurs étiquettes prosodiques. À cette fin, en tirant parti de notre analyseur de texte chinois interne, nous avons développé un analyseur de texte en langue Dungan, comme illustré dans la figure 2. Le processus commence par la normalisation et la segmentation de la phrase Dungan saisie pour déterminer la limite des mots. Une analyse des limites prosodiques suit cela pour identifier à la fois la limite du mot prosodique et celle de la phrase prosodique. Dans la dernière étape, les initiales et les finales des caractères Dungan sont dérivées grâce à un processus de conversion de caractères en unités basé sur un transformateur.

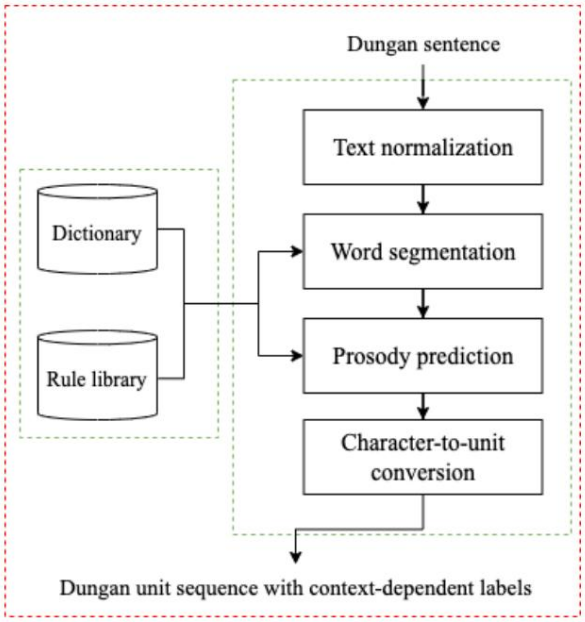


Figure 2. Procédure d'analyse du texte Dungan.

### 2.1.1. Unité de synthèse vocale de la langue Dungan

Bien qu'il utilise un système d'écriture différent, Dungan représente une prononciation dialectale du mandarin en dehors de la Chine. La langue Dungan est écrite en écriture cyrillique, ressemblant à des langues slaves comme le russe, donc la langue Dungan est composée de caractères phonétiques avec une orthographe séquentielle, suivant une structure similaire au chinois [33-35]. L'ordre orthographique des caractères Dungan se compose des initiales, des finales et du ton, comme le montre la figure

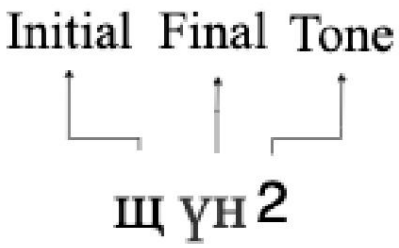


Figure 3. Structure d'un personnage Dungan.

Cet article utilise les initiales et les finales comme unité de synthèse vocale. Le personnage Dungan comprend 25 initiales (y compris l'initiale zéro) et 32 finales, comme le montre le tableau 1. Comme le mandarin, les tons de la langue Dungan sont cruciaux pour distinguer la sémantique et les émotions [36]. Dungan comporte quatre tons, à l'exclusion du ton clair, à savoir le ton de niveau (21), le ton montant (24), le ton descendant-montant (53) et le ton descendant (44), chacun désigné par les chiffres 1 à 4, respectivement.

Tableau 1. Les initiales et les finales de Dungan Language.

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

2.1.2. Normalisation du texte

Toute phrase d'entrée peut contenir des formes numériques d'heure, de date, d'abréviations et de noms propriétaires spéciaux. Avant de convertir une phrase en une séquence de symboles phonétiques, il est essentiel d'utiliser la normalisation du texte pour transformer un texte non standard en un symbole phonétique unifié. Par conséquent, nous avons implémenté une normalisation de texte basée sur des règles pour identifier les caractères non-Dungan. Nous avons développé un ensemble de règles de normalisation de texte Dungan basées sur les règles de normalisation de texte chinois [37] et avons utilisé la méthode d'ajout-restauration pour normaliser les caractères Dungan selon [38].

2.1.3. Segmentation des mots

Les limites des mots jouent un rôle important dans la prédiction des limites prosodiques. Il est donc essentiel d'identifier les limites des mots d'une phrase après normalisation. Les phrases Dungan présentent des distinctions claires entre les mots et les syllabes, ce qui rend la segmentation relativement simple. Nous avons utilisé un algorithme de segmentation de mots basé sur une correspondance maximale pour extraire les mots Dungan de la phrase d'entrée. Nous avons compilé un dictionnaire de mots Dungan comprenant 49 293 mots pour faciliter ce processus. Le mot le plus long comporte huit caractères dans ce dictionnaire, tandis que le mot le plus court comporte un seul caractère. Le dictionnaire englobe principalement les termes fondamentaux du Dungan, tels que référencés dans des sources telles que « Common Dictionary of Dungan Language » [39], « A Survey on Tungan Language in Central Asia » [40], « A Survey of Dungan Language » [41] et Termes Dungan consultables supplémentaires disponibles en ligne.

2.1.4. Prédiction des limites prosodiques

Notre approche utilise les initiales et les finales, ainsi que leurs étiquettes prosodiques, comme séquence d'unités d'entrée pour le modèle acoustique. Ainsi, extraire la structure prosodique des phrases Dun-gan est crucial pour synthétiser un discours de haute qualité. Comme le mandarin, la hiérarchie prosodique de Dungan peut être segmentée en mots prosodiques, phrases prosodiques, phrases d'intonation et pauses de phrase. La limite des phrases d'intonation peut être facilement identifiée à l'aide des signes de ponctuation Dungan. Dans cette étude, nous avons utilisé une méthode basée sur BiLSTM avec un champ aléatoire conditionnel (BiLSTM_CRF), comme illustré dans la figure 4, pour prédire les limites des mots et des phrases prosodiques [42].

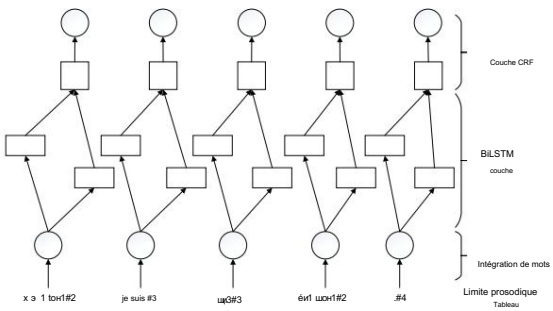


Figure 4. Le cadre de prévision des limites prosodiques Dungan basé sur BLSTM_CRF. L'entrée est une phrase Dungan avec des informations prosodiques.

Nous avons utilisé quatre ensembles d'étiquetage de position de mot prosodique distincts (#1, #2, #3, #4) pour classer les mots Dungan en phrases prosodiques. Plus précisément, #1 était utilisé pour désigner les mots prosodiques, #2 désignait les phrases prosodiques, #3 marquait la fin d'un Dungan.

mot, et le numéro 4 indiquait une pause dans une phrase. Le processus d'étiquetage incorporait des phrases et des informations prosodiques dérivées du texte Dungan, qui était étiqueté manuellement. Au cours de cette phase, les linguistes ont révisé et modifié sporadiquement des phrases sélectionnées. Nous avons atteint un haut niveau de cohérence avec les experts linguistiques grâce à des corrections itératives.

Malgré la capacité du BiLSTM à apprendre des informations dépendantes du contexte, ses décisions de classification indépendantes sont contraintes par de fortes dépendances à travers l'étiquette de sortie. Pour résoudre ce problème, nous utilisons une couche CRF qui prend en compte les balises voisines, comme illustré dans la figure 4. Pour une phrase d'entrée normalisée X = {x1, x2, · · · , xn} contenant n mots et une séquence de balises de phrase y = ( y1, y2, . , yn), chaque mot est représenté comme un vecteur de dimension D par word2vec. Nous définissons son score de prédiction s(X, y) comme suit :

$$s(X, y) = \sum_{je=1}^{n} Pi, yi + \sum_{je=0}^{n} Ayi , yi+1 \quad (1)$$

où P est la matrice des scores produits par le réseau BLSTM. Pi,yi correspond au score de la balise yi du ième mot dans une phrase. A est la matrice des scores de transition de la couche CRF, et Ayi ,yi+1 correspond au score du tag yi au tag yi+1 .

Dans la formation, nous maximisons les fonctions de log-vraisemblance suivantes :

$$\log(p(y \mid X)) = s(X, y) - \log \sum_{y \quad YX} e^{s(X,y)} \quad (2)$$

où YX représente toutes les séquences de balises possibles pour un texte d'entrée

Dans le décodage, la séquence optimale $y^*$ X. est donné comme

$$\text{oui} \quad s(X, y) = \text{suit :} \quad \text{argmax } y \quad YX \quad (3)$$

2.1.5. Conversion de caractère en unité basée sur un transformateur

Le mandarin et le Dungan utilisent le même système Pinyin pour l'étiquetage de la prononciation. Par conséquent, la conversion caractère-unité en Dungan est parallèle à celle du mandarin. Cette étude introduit une approche basée sur un transformateur [43] pour dériver l'unité Dungan, comme illustré dans la figure 5, afin d'améliorer la précision de la conversion caractère Dungan en unité. L'encodeur et le décodeur sont formés en empilant les mêmes couches essentielles avec N = 6. Chaque couche sous-jacente se compose de deux sous-couches. La première sous-couche est la couche d'attention multi-têtes. Le décodeur possède une couche d'attention multi-têtes cachée (attention multi-têtes masquée).

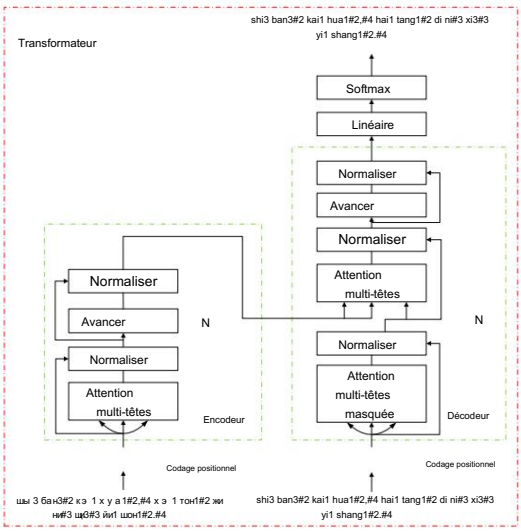


Figure 5. Le cadre de conversion de caractère Dungan en unité basé sur Transformer. L'entrée est une phrase Dungan avec des informations prosodiques (à gauche) et sa séquence Pinyin correspondante (à droite). Le résultat est la séquence Pinyin avec des informations prosodiques.

2.2. Transfert du modèle acoustique Dungan basé sur

l'apprentissage Nous implémentons le modèle acoustique Dungan en affinant un mandarin pré-entraîné modèle acoustique, comme illustré à la figure 6.

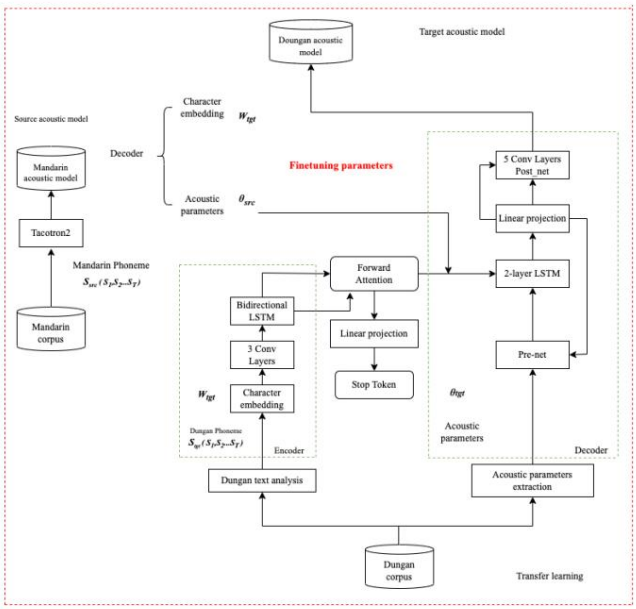


Figure 6. Procédure de formation du modèle acoustique du langage Dungan avec apprentissage par transfert.

2.3. Modèle acoustique mandarin pré-entraîné basé sur Tacotron2

Le modèle acoustique mandarin est initialement formé à l'aide d'un corpus mandarin à grande échelle. Notre analyseur de texte chinois exclusif extrait les initiales, les finales et les étiquettes prosodiques associées de ces phrases. Les caractéristiques acoustiques extraites englobent le spectrogramme Mel du corpus Mandarin à grande échelle dans le cadre Tacotron2.

Compte tenu de la prononciation similaire entre la langue Dungan et le mandarin, nous utilisons la méthode d'apprentissage par transfert par cartographie [44] pour obtenir un modèle acoustique Dungan (langue cible) en transférant des connaissances du mandarin (langue source), qui peut être formulé comme suit :

$$f\theta,W : XL \rightarrow Y \quad (4)$$

où θ représente les paramètres du modèle acoustique, W désigne les incorporations de symboles pouvant être apprises et Y représente l'espace du mandarin. XL est l'espace de texte de la langue Dungan.

$$XL = \{st\}_{t=1}^{T} \mid tst \quad L, T \quad N \quad (5)$$

où L est l'unité définie pour la langue Dungan, St est la t-ième unité de la séquence d'unités Dungan et T est la longueur de la séquence d'unités.

Dans l'encodeur, nous saisissons une séquence d'unités Dungan représentée par des intégrations de caractères. Ceci passe par une pile de trois couches convolutives, suivie d' une normalisation par lots et d'activations ReLU. Par la suite, la sortie de la couche convolutive finale est introduite dans une couche LSTM bidirectionnelle pour générer les caractéristiques de l'unité Dungan.

L'apprentissage par transfert basé sur la cartographie implique la cartographie des instances de θsrc et θtgt dans un nouvel espace de paramètres acoustiques. Dans ce processus, nous pouvons utiliser directement Wsrc et θsrc décodés du modèle acoustique mandarin par le décodeur. θsrc et θtgt peuvent prendre des intégrations en entrée et générer de la parole. Cependant, comme ssrc et stgt proviennent de jeux de symboles différents, c'est-à-dire Lsrc = Ltgt, le même concept ne peut pas être appliqué directement à Wsrc et Wtgt. Pour résoudre ce problème, les unités Dunggan sont intégrées dans Wtgt pour faciliter le réapprentissage pendant le processus de transmission.

Nous adoptons le mécanisme d'attention directe, qui utilise des pondérations d'attention cumulatives. pour calculer le vecteur de contexte.

Le décodeur est un réseau neuronal récurrent autorégressif qui prédit un θtgt à partir de l'unité Dungan d'entrée du codeur séquence une image à la fois. Nous pouvons utiliser θsrc appris du modèle acoustique mandarin pour initialiser θtgt dans le nouvel espace de paramètres acoustiques. La sortie du pas de temps initial est d'abord traitée via un pré-réseau composé de deux couches entièrement connectées. Cette sortie est combinée avec le contexte d'attention avancée vecteur et passé à travers une paire de couches LSTM. La combinaison des sorties LSTM et les vecteurs de contexte d'attention subissent trois transformations linéaires distinctes pour prédire le trame de spectrogramme cible, jeton d'arrêt et résidu estimé. Par la suite, la prévision les caractéristiques acoustiques sont soumises à cinq couches convolutives, générant un résidu pour améliorer la reconstruction du modèle acoustique Dungan.

## 3. Résultats

### 3.1. Évaluation de la conversion de personnage en unité Dungan à base de transformateur

L'analyse du texte au front-end affecte la qualité de la synthèse vocale au back-end fin, nous avons donc évalué l'analyseur de texte Dungan, où la conversion caractère en unité Le module est le facteur le plus critique affectant la qualité de la parole synthétisée. Pour évaluer la viabilité du module de conversion caractère-unité Dungan basé sur un transformateur, nous utilisé un ensemble de données comprenant 10 783 phrases en langue Dungan transcrites en utilisant Mandarine Pinyin. Langue Dungan de l'ensemble de données et représentations en mandarin pinyin sont isomorphes, encapsulant des attributs textuels comme le ton et les limites prosodiques inhérentes à la langue Dungan. Dans notre recherche, nous avons alloué 10 % du total de 10 783 phrases pour servir d'ensemble de test, 10 % supplémentaires comme ensemble de validation, et les 80 % restants ont été désigné comme ensemble de formation. Les hyperparamètres associés au Transformer sont détaillé dans le tableau 2. Nous avons utilisé des mesures de précision, de rappel et de F1 comme indices d'évaluation, ainsi que illustré dans le tableau 3. Les résultats du processus d'évaluation ont confirmé que le projet proposé Le module caractère-unité Dungan convient à une évaluation ultérieure de la synthèse vocale.

Tableau 2. Les hyperparamètres du modèle de conversion de caractère en unité basé sur un transformateur.

| Paramètre | Valeur |
|---|---|
| Couches d'attention Nx | 6 |
| Têtes | 8 |
| Taille du lot | 32 |
| Caché | 513 |
| Abandonner | 0,1 |
| Taux d'apprentissage | 0,0001 |

Tableau 3. Résultats de la conversion caractère-unité Dungan basée sur un transformateur.

| Précision | Rappel | F1 |
|---|---|---|
| 90.12 | 89.91 | 90.01 |

### 3.2. Évaluation des modèles acoustiques Dungan basés sur l'apprentissage par transfert

#### 3.2.1. Corpus

Dans l'expérience, nous avons utilisé les enregistrements de neuf locuteurs féminins et de trente et un hommes. de la base de données chinoise Tsinghua de 30 heures [45] (totalisant 13 389 phrases) comme le corpus mandarin. Pour le corpus Dungan, nous avons sélectionné cinq enregistrements de locuteurs masculins (923 par personne, totalisant 4615 phrases et 6 h). Le corpus Dungan englobe tous les éléments initiaux et prononciations finales de la langue Dungan. La longueur moyenne des phrases est de 18 syllabes, avec une durée moyenne de 10 s. Tous les enregistrements ont été convertis en monocanal 16 kHz fréquence d'échantillonnage avec une précision de quantification de 16 bits.

3.2.2. Montage expérimental

Trois types de frameworks TTS, dont Tacotron+Griffin-Lim, Tacotron2+WaveNet, et Tacotron2+WaveRNN, ont été comparés dans les expériences. Quelques hyperparamètres de les cadres sont fournis dans le tableau 4.

Tableau 4. Hyperparamètres du modèle de Tacotron et Tacotron2.

| Modèle | | Tacotron | Tacotron2 | Tacotron2 au garde-à-vous |
|---|---|---|---|---|
| Vocodeur | | Griffin-Lim | WaveNet | OndeRNN |
| Encodeur | Intégration | Phomème (256) | Phomème (512) | Phomème (512) |
| | Pré-net | FFN (256, 128) | - | FFN Phomème (512, 256) |
| | Noyau d'encodeur | CBHG (256) | CNN (512)<br>Bi-LSTM (512) | CNN (256)<br>Bi-LSTM (256, 512) |
| Décodeur | Post-net | CBHG (256) | CNN (512) | CNN (512) |
| | Décodeur RNN | GRU (256, 256) | - | LSTM (512, 256) |
| | Attention | Additif (256) | Dépend de l'emplacement (128) | En avant (256) |
| | Attention RNN | GRU (256) | LSTM (1024, 1024) | LSTM (256) |
| | Pré-net | FFN (256, 128) | FFN (256, 256) | FFN (256, 128) |
| Paramètre | | 7,6 × 106 | 28,9 × 106 | 23,7 × 106 |

Les trois frameworks comprennent un module d'analyse de texte frontal, un modèle acoustique module de formation et un vocodeur. Le module analyseur de texte transforme le doungan ou le chinois phrases dans une séquence d'unités représentée en pinyin, y compris les initiales, les finales et leurs tons et les étiquettes de limites prosodiques. Dans le module de formation du modèle acoustique, nous dérivons le journal spectrogramme de magnitude à partir du signal vocal en utilisant le fenêtrage de Hann avec une fréquence de 80 ms longueur de trame, décalage de trame de 12,5 ms et transformation de Fourier de 2048 points.

Pour le framework Tacotron+Griffin-Lim, les modèles acoustiques sont formés à l'aide d'une sortie facteur de réduction de couche de r = 3 et l'optimiseur Adam avec un taux d'apprentissage décroissant. Le le taux d'apprentissage commence à 0,001 et est ensuite réduit à 0,0005, 0,0003 et 0,0001 après 5, 20 et 50 époques, respectivement. Une fonction de perte simple est utilisée pour le décodeur seq2seq (spectrogramme Mel) et réseau de post-traitement (spectrogramme linéaire). La taille du lot d'entraînement est définie sur 32, toutes les séquences étant complétées jusqu'à une longueur maximale de reconstruire les trames avec remplissage de zéros. L'algorithme Griffin-Lim est utilisé comme vocodeur pour la conversion du spectre en parole de Mel.

Pour le framework basé sur Tacotron2+WaveNet, nous entraînons les modèles acoustiques à l'aide du procédure standard de formation à maximum de vraisemblance, qui consiste à fournir le résultat correct au lieu de la sortie prévue du côté du décodeur. Ceci a été complété par une taille de lot sur 32. L'optimiseur Adam a été utilisé avec les paramètres définis comme suit : $\beta = 0{,}9$, $\beta = 0{,}999$, $= 10-6$ . Le taux d'apprentissage a été initialisé à $10-3$ puis a diminué de façon exponentielle jusqu'à $10-5$ après 50 000. De plus, nous avons appliqué la régularisation L2 avec un poids de $10-6$ . Pour le Mel conversion spectre-parole, le WaveNet a été utilisé comme vocodeur.

Dans notre cadre d'apprentissage par transfert basé sur Tacotron2+WaveRNN, nous employons initialement un Corpus mandarin à grande échelle pour pré-entraîner un modèle acoustique mandarin pour le modèle suivant transfert. Ce modèle pré-entraîné est ensuite utilisé pour entraîner le modèle acoustique Dungan via transfert apprendre du corpus mandarin-dungan. Pour le vocodage, nous utilisons le WaveRNN pour Conversion du spectre en parole. Étant donné que les réglages des paramètres ont un impact significatif précision et robustesse du modèle, nous avons optimisé ces paramètres grâce à un entraînement itératif et mises à jour.

Chaque framework TTS implémente une synthèse vocale monolingue pour le mandarin ou le Dungan et une synthèse vocale bilingue basée sur l'apprentissage par transfert. Nous avons formé plusieurs modèles sur trois

Cadres TTS pour évaluer la qualité et la clarté de la parole synthétisée. Dans notre expérience, 10 % des énoncés ont été attribués au hasard à l'ensemble de test, 10 % supplémentaires ont été désignés pour l'ensemble de développement, et les énoncés restants constituaient l'ensemble d'apprentissage.

Modèle Dungan monolingue dépendant du locuteur

Nous avons formé le modèle acoustique Dungan Monolingual Speaker-Dependent (DSD) en utilisant des enregistrements de cinq locuteurs masculins, chacun contribuant à 923 phrases, totalisant 4615. phrases et s'étalant sur 6 h. Nous avons ensuite comparé la qualité et la clarté des signaux synthétisés discours à travers trois frameworks : DSD-Tacotron+Griffin-Lim, DSD Tacotron2+WaveNet, et DSD-Tacotron2+WaveRNN.

Modèle dépendant du locuteur monolingue mandarin

Nous avons utilisé des enregistrements de neuf locuteurs féminins et de trente et un hommes (Tsinghua Base de données chinoise de 30 heures, composée de 13 389 phrases) pour entraîner le mandarin monolingue Modèle acoustique dépendant du haut-parleur (MSD). Nous avons comparé la qualité de la parole synthétisée et clarté dans trois frameworks : MSD-Tacotron+Griffin-Lim, MSD-Tacotron2+WaveNet, et MSD-Tacotron2+WaveRNN.

Modèle dépendant du locuteur bilingue mandarin et dungan

Nous avons utilisé des enregistrements de cinq locuteurs masculins Dungan (923 phrases par individu, totalisant jusqu'à 4615 phrases, équivalentes à 6 h) comme données d'entraînement pour transférer le modèle acoustique mandarin au modèle acoustique Dungan afin de réaliser un modèle dépendant du locuteur Dungan (MDSD) et un modèle acoustique mandarin dépendant du locuteur (MDSM). Nous puis comparé la qualité et la clarté de la parole synthétisée dans six cadres.

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

### 3.2.3. Évaluations objectives

Nous avons utilisé la distorsion Mel-cepstrale (MCD) [46], Band A Periodicity Distortion (BAP) [47], erreur quadratique moyenne (RMSE) [48] et erreur voisée/non voisée (V/UV) [47] évaluer objectivement les différents modèles. Les résultats pour le DSD et le MSD acoustique les modèles sont présentés respectivement dans le tableau 5 et le tableau 6. De même, le MDSM et le MDSD les résultats sont affichés dans le tableau 7 et le tableau 8, respectivement.

Tableau 5. Résultats objectifs du modèle acoustique DSD pour Dungan.

| Modèle | Tacotron+Griffin-Lim | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| MDC (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 |

Tableau 6. Résultats objectifs du modèle acoustique MSD pour le mandarin.

| Modèle | Tacotron+Griffin-Lim | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| MDC (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 |

Tableau 7. Résultats objectifs du modèle acoustique MDSD pour Dungan.

| Modèle | Tacotron+Griffin-Lim | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| MDC (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 |

Tableau 8. Résultats objectifs du modèle acoustique MDSM pour le mandarin.

| Modèle | 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 |

Dans le contexte de la synthèse vocale Dungan à faibles ressources, la qualité de l'alignement de l'attention entre l'encodeur et le décodeur influence de manière significative la qualité de la parole synthétisée. discours. Les désalignements sont principalement évidents au niveau de la lisibilité, des sauts et des répétitions. Par conséquent, nous utilisons le taux de mise au point diagonale (DFR) et le taux d'intelligibilité au niveau des mots. (IR) [49] pour évaluer la lisibilité dans les langues à faibles ressources, comme illustré dans le tableau 9. Le DFR représente la carte d'attention entre l'encodeur et le décodeur, servant de carte architecturale métrique. L'IR mesure le pourcentage de mots tests prononcés correctement et clairement par les humains, une mesure standard pour évaluer la qualité de la génération vocale à faibles ressources.

Tableau 9. Lisibilité du discours Dungan synthétisé.

| Modèle | RI (%) | RFFA (%) |
|---|---|---|
| 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. Évaluation subjective

Pour les évaluations subjectives, 30 phrases ont été sélectionnées au hasard dans l'ensemble de tests. Nous avons réalisé trois tests : score d'opinion moyen (MOS), score d'opinion moyen de dégradation (DMOS) et préférence AB pour évaluer la qualité de la parole synthétisée. Nous avons recruté 20 locuteurs natifs du mandarin et 10 étudiants internationaux natifs Dungan (qui comprenaient chinois) comme participants. Ces participants ont reçu une formation avant l'évaluation formelle. Les participants mandarin ont évalué les modèles acoustiques mandarin de MSD et MDSM, tandis que les participants de Dungan ont évalué les modèles acoustiques Dungan du DSD et du MDSD. Au cours du test MOS, les participants ont évalué le naturel de la parole synthétisée sur une note de 5 points. échelle. Les scores MOS moyens pour les discours synthétisés en Dungan et en Mandarin sont présentés dans les figures 7 et 8.

Dans le test DMOS, l'énoncé synthétisé de chaque modèle et l'original correspondant l'enregistrement comprenait une paire de fichiers vocaux. Ces paires ont été jouées au hasard sujets, le discours synthétisé précédant l'original. Les participants ont été chargés en comparant méticuleusement les deux fichiers et en évaluant la similitude des résultats synthétisés. discours à l'original sur une échelle de 5 points. Un score de 5 indique que la synthèse le discours était similaire à l'original, alors qu'un score de 1 signifiait une disparité significative. Les figures 9 et 10 montrent les scores DMOS moyens pour le Dungan et le Mandarin synthétisés. discours, respectivement.

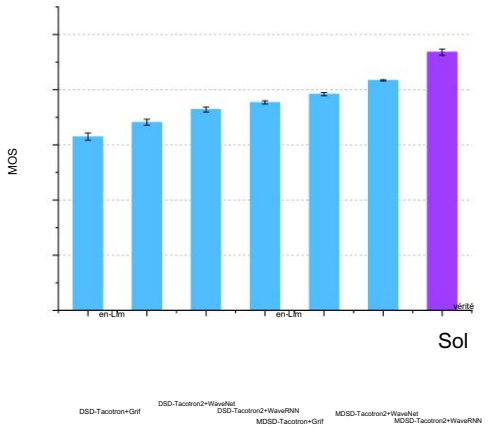


Figure 7. Les scores MOS moyens du discours Dungan synthétisé sous des intervalles de confiance de 95 %.

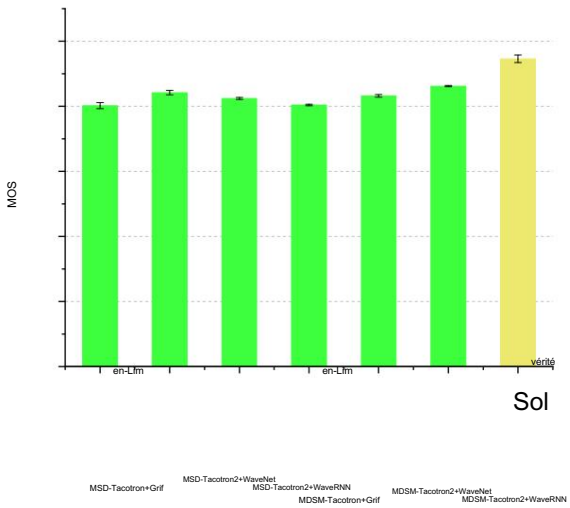


Figure 8. Les scores MOS moyens du discours synthétisé en mandarin sous des intervalles de confiance de 95 %.

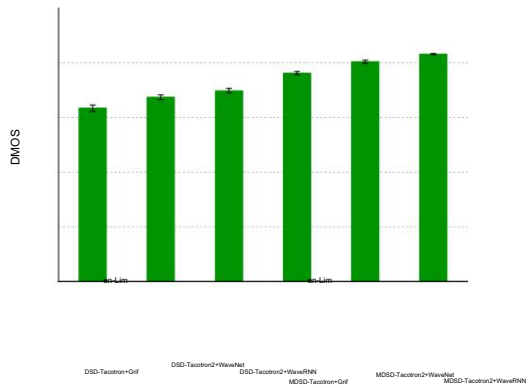


Figure 9. Les scores DMOS moyens du discours Dungan synthétisé sous des intervalles de confiance de 95 %.

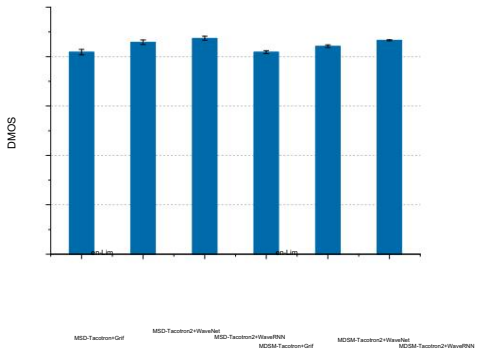


Figure 10. Les scores DMOS moyens du discours synthétisé en mandarin sous des intervalles de confiance de 95 %.

Dans le test de préférence AB, chaque paire était composée de deux phrases identiques. Les énoncés synthétisés ont été joués dans un ordre aléatoire. Les participants ont été invités à écouter et évaluez quel énoncé avait une qualité supérieure ou indiquez « neutre » si aucune préférence a été discerné. Les résultats synthétisés des préférences vocales en Dungan et en Mandarin sont présentés dans les tableaux 10 et 11, respectivement.

Tableau 10. Score de préférence AB subjectif (%) de Dungan avec ρ < 0,01.

| | DSD-Tacotron+ Griffin-Lim | DSD-Tacotron2+ WaveNet | DSD-Tacotron2 +WaveRNN | MDSD-Tacotron+ Griffin-Lim | MDSD-Tacotron2 +WaveNet | MDSD-Tacotron2+ OndeRNN | Neutre |
|---|---|---|---|---|---|---|---|
| 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 |

Tableau 11. Score de préférence AB subjectif (%) du mandarin avec ρ < 0,01.

| | MSD-Tacotron+ Griffin-Lim | MSD-Tacotron2+ WaveNet | MSD-Tacotron2+ OndeRNN | MDSM-Tacotron+ Griffin-Lim | MDSM-Tacotron2+ WaveNet | MDSM-Tacotron2+ OndeRNN | Neutre |
|---|---|---|---|---|---|---|---|
| 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

Dans les évaluations objectives, bien que le cadre TTS basé sur Tacotron+Griffin-Lim mappe les caractéristiques linguistiques aux caractéristiques acoustiques image par image à travers le corpus monolingue Dungan , le discours Dungan synthétisé doit améliorer sa qualité et sa lisibilité. Cependant, une attention particulière et un modèle acoustique affiné peuvent améliorer la lisibilité. et réduire le temps de formation. Par conséquent, le Tacotron2+WaveRNN basé sur l'apprentissage par transfert Le modèle acoustique du framework surpasse les autres. Les résultats objectifs du modèle acoustique MDSD dépassent ceux du modèle acoustique DSD. C'est parce que Dungan est une variante du dialecte du nord-ouest de la Chine, qui partage de nombreuses similitudes internes. Étant donné les similitudes de prononciation entre le mandarin et le Dungan, le même symbole représente leur prononciations exactes. Par conséquent, nous concluons que l'ajout d'un corpus en mandarin et l'utilisation l'apprentissage par transfert peut améliorer la qualité et la lisibilité du discours Dungan synthétisé.

Toutes les évaluations subjectives s'alignent sur les évaluations objectives sous divers aspects. Le Le cadre Tacotron2+waveRNN basé sur l'apprentissage par transfert offre une qualité vocale supérieure,

notamment en ce qui concerne le naturel et la lisibilité de la parole synthétisée. Avec l' ajout du corpus mandarin, la qualité et la lisibilité du discours Dungan synthétisé à l'aide des cadres TTS basés sur l'apprentissage par transfert surpassent ceux des cadres TTS formés par corpus monolingue. Ceci est en outre validé par le test de préférence AB, qui confirme que nos cadres TTS proposés offrent une qualité et une lisibilité améliorées par rapport à la parole synthétisée par le modèle acoustique monolingue.

## 5. Conclusions

Cette étude étend nos recherches antérieures en mettant en œuvre une synthèse vocale mandarin basée sur l'apprentissage par transfert et une synthèse vocale Dungan à faibles ressources dans le cadre Tacotron2+WaveRNN. Nous avons également développé un analyseur de texte Dungan complet. Des expériences objectives et subjectives ont révélé que la synthèse vocale Dungan basée sur l'apprentissage par transfert dans le cadre Tacotron2+WaveRNN a surpassé les méthodes alternatives et le cadre de synthèse vocale monolingue Dungan. De plus, l'apprentissage par transfert n'a pas compromis la qualité vocale et la lisibilité du discours Dungan synthétisé à faibles ressources. Par conséquent, notre approche présente un potentiel important pour développer des systèmes de synthèse vocale pour les langues minoritaires à faibles ressources.

De nombreuses avancées ont été réalisées dans le domaine du TTS basé sur des réseaux neuronaux profonds. Nous avons remarqué que de nouvelles méthodes de synthèse vocale [50–52] ont été proposées récemment. Motivées par les progrès récents dans les modèles auto-régressifs (AR) employant des architectures de décodeur uniquement pour la génération de texte, plusieurs études, telles que VALL-E [53] et BASE TTS [54], appliquent des architectures similaires aux tâches TTS. Ces études démontrent la capacité remarquable des architectures composées uniquement de décodeurs à produire une parole au son natu Ces études démontrent la capacité remarquable des architectures composées uniquement de décodeurs à produire une parole au son naturel. Les recherches futures se concentreront sur l'utilisation de ces nouvelles méthodes pour améliorer la qualité de la synthèse vocale des langues Dungan, réduire la taille du corpus Dungan et réaliser une synthèse vocale pour les langues Dungan en utilisant un corpus plus grand. De plus, l'apprentissage multitâche sera exploré pour réaliser des scénarios indépendants du locuteur et améliorer l'émotion du discours Dungan synthétisé.

Contributions de l'auteur : Conceptualisation, ML et HY ; analyse formelle, HY et RJ ; conservation des données , ML et RJ ; rédaction – préparation de l'ébauche originale, ML et RJ ; rédaction – révision et édition, HY et ML ; supervision, HY; acquisition de financement, HY Tous les auteurs ont lu et accepté la version publiée du manuscrit.

Financement : La recherche est soutenue par le fonds de recherche de la Fondation nationale des sciences naturelles de Chine (subvention n° 62067008).

Déclaration de l'Institutional Review Board : Ne s'applique pas aux études n'impliquant pas des humains ou des animaux.

Déclaration de consentement éclairé : sans objet.

Déclaration de disponibilité des données : nous avons utilisé deux ensembles de données de formation dans le manuscrit. L'un est un ensemble de données en mandarin accessible au public (THCHS-30) et l'autre est un ensemble de données Donggan, comprenant la parole et le texte. Le premier est public et peut être consulté à partir de http://www.openslr.org/18/ (consulté le 16 juin 2024). Ce dernier est un ensemble de données auto-construit et n'est pas accessible au public. Toutefois, les données seront mises à disposition sur demande.

Conflits d'intérêts : Les auteurs ne déclarent aucun conflit d'intérêts. Les bailleurs de fonds n'ont joué aucun rôle dans la conception de l'étude ; dans la collecte, l'analyse ou l'interprétation des données ; dans la rédaction du manuscrit ; ou dans la décision de publier les résultats.

## Les références

1. Tu, T. ; Chen, YJ; Chieh Yeh, C. ; Yi Lee, H. Texte-parole de bout en bout pour les langues à faibles ressources par transfert interlingue Apprentissage. arXiv 2019, arXiv :1904.06508.
2. Liu, R. ; Sisman, B. ; Bao, F. ; Yang, J. ; Gao, G. ; Li, H. Exploiter les caractéristiques morphologiques et phonologiques pour améliorer le phrasé prosodique pour la synthèse vocale mongole. Trans. IEEE/ACM. Langage vocal audio. Processus. 2021, 29, 274-285. [Référence croisée]
3. Saeki, T. ; Maiti, S. ; Li, X. ; Watanabe, S. ; Takamichi, S. ; Saruwatari, H. Adaptation du langage basée sur le graphone inductif pour la synthèse vocale à faibles ressources. Trans. IEEE/ACM. Langage vocal audio. Processus. 2024, 32, 1829-1844. [Référence croisée]

4. Xu, J. ; Tan, X. ; Ren, Y. ; Qin, T. ; Li, J. ; Zhao, S. ; Liu, TY LRSpeech : synthèse et reconnaissance vocale à ressources extrêmement faibles. Dans les actes de la 26e conférence internationale ACM SIGKDD sur la découverte des connaissances et l'exploration de données, KDD'20, New York, NY, États-Unis, 6-10 juillet 2020 ; pages 2802 à 2812.
[Référence croisée]
5. Lui, M. ; Yang, J. ; Lui, L. ; Soong, FK Modèles multilingues Byte2Speech pour une synthèse vocale évolutive à faibles ressources. arXiv 2021, arXiv :2103.03541.
6. Oliveira, FS; Casanova, E. ; Junior, CA ; Soares, AS; Galvão Filho, AR CML-TTS : un ensemble de données multilingues pour la synthèse vocale dans les langues à faibles ressources. Dans le texte, la parole et le dialogue ; Ekštein, K., Pártl, F., Konopík, M., éd.; Springer : Cham, Suisse, 2023 ; pp. 188-199.
7. Zhu, Y. Langue Donggan : une variété spéciale des dialectes du Shaanxi et du Gansu. Langue asiatique. Culte. 2013, 4, 51-60. 8.
Jiang, Y. Langue Donggan et sa relation avec les dialectes du Shaanxi et du Gansu. J. Chin. Linguiste. 2014, 42, 229-258.
9. Chen, L. ; Yang, H. ; Wang, H. Recherche sur la synthèse vocale Dungan basée sur Deep Neural Network. Dans Actes du 11e Symposium international 2018 sur le traitement de la langue parlée chinoise (ISCSLP), Taipei, Taiwan, 26-29 novembre 2018 ; p. 46-50. [Référence croisée]
10. Jiang, R. ; Chen, C. ; Shan, X. ; Yang, H. Utilisation de l'amélioration de la parole pour réaliser la synthèse vocale des langues Dungan à faibles ressources. Dans Actes de la 24e Conférence 2021 du Comité international oriental de la COCOSDA pour la coordination et la normalisation des bases de données vocales et des techniques d'évaluation (O-COCOSDA), Singapour, 18-20 novembre 2021 ; pp. 193-198. [Référence croisée]
11. Chasse, AJ ; Noir, sélection d'unité AW dans un système de synthèse vocale concaténative utilisant une grande base de données vocale. Dans les actes de la conférence internationale IEEE de 1996 sur l'acoustique, la parole et le traitement du signal, Atlanta, GA, États-Unis, 9 mai 1996 ; Volume 1, p. 373-376.
12. Tokuda, K. ; Nankaku, Y. ; Aujourd'hui, T. ; Zen, H. ; Yamagishi, J. ; Oura, K. Synthèse vocale basée sur des modèles de Markov cachés. Proc. IEEE 2013, 101, 1234-1252. [Référence croisée]
13. Ling, ZH; Deng, L. ; Yu, D. Modélisation d'enveloppes spectrales à l'aide de machines Boltzmann restreintes et de réseaux de croyances profondes pour la synthèse vocale paramétrique statistique. IEEETrans. Langage vocal audio. Processus. 2013, 21, 2129-2139. [Référence croisée]
14. Zen, H. ; Aîné, A. ; Schuster, M. Synthèse vocale paramétrique statistique utilisant des réseaux de neurones profonds. Dans les actes de la conférence internationale IEEE 2013 sur l'acoustique, la parole et le traitement du signal, Vancouver, Colombie-Britannique, Canada, 26-31 mai 2013 ; pages 7962 à 7966. [Référence croisée]
15. Wang, P. ; Qian, Y. ; Bientôt, FK ; Lui, L. ; Zhao, H. Intégration de mots pour la synthèse TTS récurrente basée sur un réseau neuronal. Dans les actes de la conférence internationale IEEE 2015 sur l'acoustique, la parole et le traitement du signal (ICASSP), South Brisbane, QLD, Australie, 19-24 avril 2015 ; pages 4879 à 4883. [Référence croisée]
16. Yu, Q. ; Liu, P. ; Wu, Z. ; Ang, Saskatchewan ; Meng, H. ; Cai, L. Apprentissage d'informations multilingues avec BLSTM multilingue pour la synthèse vocale de langues à faibles ressources. Dans les actes de la conférence internationale IEEE 2016 sur l'acoustique, la parole et le traitement du signal (ICASSP), Shanghai, Chine, 20-25 mars 2016 ; pages 5545 à 5549. [Référence croisée]
17. Tan, X. ; Chen, J. ; Liu, H. ; Cong, J. ; Zhang, C. ; Liu, Y. ; Wang, X. ; Leng, Y. ; Yi, Y. ; Lui, L. ; et coll. NaturalSpeech : synthèse texte-parole de bout en bout avec une qualité de niveau humain. IEEETrans. Modèle Anal. Mach. Intell. 2024, 46, 4234-4245. [Référence croisée]
18. Wang, Y. ; Skerry-Ryan, RJ; Stanton, D. ; Wu, Y. ; Weiss, RJ; Jaitly, N. ; Yang, Z. ; Xiao, Y. ; Chen, Z. ; Bengio, S. ; et coll. Tacotron : vers une synthèse vocale de bout en bout. Dans Actes de la 18e conférence annuelle de l'International Speech Communication Association, Interspeech 2017, Stockholm, Suède, 20-24 août 2017.
19. Shen, J. ; Pang, R. ; Weiss, RJ; Schuster, M. ; Jaitly, N. ; Yang, Z. ; Chen, Z. ; Zhang, Y. ; Wang, Y. ; Skerrv-Ryan, R. ; et coll. Synthèse TTS naturelle en conditionnant Wavenet sur les prédictions du spectrogramme MEL. Dans les actes de la conférence internationale IEEE 2018 sur l'acoustique, la parole et le traitement du signal (ICASSP), Calgary, AB, Canada, 15-20 avril 2018 ; pages 4779 à 4783. [Référence croisée]
20. Griffin, D. ; Lim, J. Estimation du signal à partir de la transformée de Fourier à court terme modifiée. IEEETrans. Acoustique. Processus du signal vocal. 1984, 32, 236-243. [Référence croisée]
21. van den Oord, A. ; Dieleman, S. ; Zen, H. ; Simonyan, K. ; Vinyals, O. ; Graves, A. ; Kalchbrenner, N. ; Aîné, A. ; Kavukcuoglu, K.
WaveNet : un modèle génératif pour l'audio brut. 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. Synthèse audio neuronale efficace. arXiv 2018, arXiv : 1802.08435.
23. Byambadorj, Z. ; Nishimura, R. ; Ayush, A. ; Ohta, K. ; Kitaoka, N. Système de synthèse vocale pour les langues à faibles ressources utilisant l'apprentissage par transfert multilingue et l'augmentation des données. EURASIP J. Musique vocale audio. Processus. 2021, 2021, 42. [Réf. croisée]
24. Joshi, R. ; Garera, N. Adaptation rapide du locuteur dans les systèmes de synthèse vocale à faibles ressources utilisant des données synthétiques et l'apprentissage par transfert. Dans Actes de la 37e Conférence de l'Asie-Pacifique sur la langue, l'information et le calcul, Hong Kong, Chine, 2-4 décembre 2023 ; Huang, CR, Harada, Y., Kim, JB, Chen, S., Hsu, YY, Chersoni, EAP, Zeng, WH, Peng, B., Li, Y. et al., Eds. ; ACL : Hong Kong, Chine, 2023 ; pp. 267-273.
25. Faites, P. ; Coler, M. ; Dijkstra, J. ; Klabbers, E. Stratégies d'apprentissage par transfert pour la synthèse vocale à faibles ressources : cartographie téléphonique, saisie des fonctionnalités et sélection de la langue source. Dans Actes du 12e atelier de synthèse vocale ISCA (SSW2023), Grenoble, France, 26-28 août 2023 ; p. 21-26. [Référence croisée]

26. Azizah, K. ; Jatmiko, W. Apprentissage par transfert, contrôle du style et perte de reconstruction du locuteur pour un multi-locuteur multilingue zéro tir synthèse vocale dans les langues à faibles ressources. Accès IEEE 2022, 10, 5895-5911. [Référence croisée]
27. Cai, Z. ; Yang, Y. ; Li, M. Synthèse vocale multilingue multi-locuteurs avec des données de formation bilingues limitées. Calculer. Langage du discours. 2023, 77, 101427. [Réf. croisée]
28. Yang, H. ; Oura, K. ; Wang, H. ; Gan, Z. ; Tokuda, K. Utilisation de la formation adaptative du locuteur pour réaliser des langues multilingues mandarin-tibétain synthèse de discours. Multimed. Outils Appl. 2015, 74, 9927-9942. [Référence croisée]
29. Wang, L. ; Yang, H. Méthode de segmentation de mots tibétains basée sur le modèle bilstm_crf. Dans les actes de la conférence internationale IEEE 2018 sur le traitement des langues asiatiques (IALP), Bandung, Indonésie, 15-17 novembre 2018 ; pp. 297-302.
30. Zhang, W. ; Yang, H. ; Bu, X. ; Wang, L. Apprentissage profond pour la synthèse vocale multilingue mandarin-tibétain. Accès IEEE 2019, 7, 167884-167894. [Référence croisée]
31. Zhang, W. ; Yang, H. Améliorer la synthèse vocale tibétaine séquence à séquence avec des informations prosodiques. ACM Trans. Asiatique à faibles ressources. Lang. Inf. Processus. 2023, 22, 6012. [Réf. croisée]
32. Zhang, W. ; Yang, H. Méta-apprentissage pour la synthèse vocale multilingue mandarin-tibétain. Appl. Sci. 2022, 12, 2185. [CrossRef]
33. Hai, F. Une étude pilote sur les mots d'emprunt dans la langue Dungan d'Asie centrale. Université du Xinjiang. J. 2000, 28, 58-63.
34. Lin, T. Caractéristiques, situation et tendances de développement de la langue toung'gan en Asie centrale. Contemp. Linguiste. 2016, 18, 234-243.
35. Gladney, DC Altérité relationnelle : construction des identités dungan (hui), ouïghoure et kazakhe à travers la Chine, l'Asie centrale et la Turquie. Hist. Anthropol. 1996, 9, 445-477. [Référence croisée]
36. Miao, DX Modèle d'enseignement bilingue du peuple Donggan. J. Rés. Éduc. Ethn. Mineure. 2008, 19, 111-114.
37. Jia, Y. ; Huang, D. ; Liu, W. ; Dong, Y. ; Yu, S. ; Wang, H. Normalisation du texte dans le système de synthèse vocale en mandarin. Dans les actes de la conférence internationale IEEE 2008 sur l'acoustique, la parole et le traitement du signal, Las Vegas, NV, États-Unis, 31 mars-4 avril 2008 ; pp. 4693-4696. [Référence croisée]
38. Wanmezhaxi, N. Recherche sur plusieurs questions clés dans la segmentation des mots tibétains. J. Chin. Inf. Processus. 2014, 28, 132-139.
39. Zavyalova, O. Langue Dungan. 2015. Disponible en ligne : https://www.academia.edu/42869092/Dungan_Language (accédé le 16 juin 2024).
40. Lin, T. Donggan Writing—Un essai réussi d'écriture alphabétique chinoise. J. Deuxièmement. Université du Nord-Ouest. Natl. 2005, 2005, 31-36.
41. Yang, WJ; Zhang, R. Identité ethnique dans le contexte transnational – Un cas d'étude sur « Dunggan » et la nationalité Hui. J. Sud-Cent. Univ. Natl. 2009, 29, 31-36.
42. Zheng, Y. ; Tao, J. ; Wen, Z. ; Li, Y. Prédiction des limites prosodiques de bout en bout basée sur BLSTM-CRF avec intégrations sensibles au contexte dans un frontal de synthèse vocale. Proc. Interdiscours 2018, 9, 47-51. [Référence croisée]
43. Hlaing, AM; Pa, WP Modèles séquence à séquence pour la conversion de graphème en phonème sur un grand dictionnaire de prononciation du Myanmar. Dans Actes de la 22e Conférence 2019 du Comité international oriental de la COCOSDA pour la coordination et la normalisation des bases de données vocales et des techniques d'évaluation (O-COCOSDA), Cebu, Philippines, 25-27 octobre 2019 ; p. 1 à 5. [Référence croisée]
44. Tan, C. ; Soleil, F. ; Kong, T. ; Zhang, W. ; Yang, C. ; Liu, C. Une enquête sur l'apprentissage par transfert profond. Dans Réseaux de neurones artificiels et apprentissage automatique — ICANN 2018 ; K ˚urková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I., Eds.; Springer : Cham, Suisse, 2018 ; pp. 270-279.
45. Wang, D. ; Zhang, X. THCHS-30 : Un corpus de parole chinoise libre. arXiv 2015, arXiv : 1512.01882.
46. Kubichek, R. Mesure de distance Mel-cepstrale pour l'évaluation objective de la qualité de la parole. Dans Actes de la conférence IEEE Pacific Rim sur les ordinateurs de communication et le traitement du signal, Victoria, Colombie-Britannique, Canada, 19-21 mai 1993 ; Volume 1, p. 125-128. [Référence croisée]
47. Dhiman, JK ; Seelamantula, CS Une technique spectro-temporelle pour estimer l'apériodicité et les limites de décision vocales/non vocales des signaux vocaux. Dans les actes de la conférence internationale IEEE 2019 sur l'acoustique, la parole et le traitement du signal (ICASSP2019), Brighton, Royaume-Uni, 12-17 mai 2019 ; pages 6510 à 6514. [Référence croisée]
48. Castelazo, I. ; Mitani, Y. Sur l'utilisation de l'erreur quadratique moyenne comme indice de compétence. Accréditer. Qual. Assurer. 2012, 17, 95-97. [Référence croisée]
49. Ren, Y. ; Tan, X. ; Qin, T. ; Zhao, S. ; Zhao, Z. ; Liu, TY Synthèse vocale presque non supervisée et reconnaissance vocale automatique. arXiv 2020, arXiv :1905.06791.
50. Ren, Y. ; Hu, C. ; Tan, X. ; Qin, T. ; Zhao, S. ; Zhao, Z. ; Liu, TY FastSpeech 2 : synthèse vocale de bout en bout rapide et de haute qualité. arXiv 2022, arXiv :2006.04558.
51. Chen, J. ; Chanson, X. ; Peng, Z. ; Zhang, B. ; Poêle, F. ; Wu, Z. LightGrad : Modèle probabiliste de diffusion léger pour la synthèse vocale. Dans les actes de la conférence internationale IEEE 2023 sur l'acoustique, la parole et le traitement du signal (ICASSP2023), île de Rhodes, Grèce, 4-10 juin 2023 ; p. 1 à 5. [Référence croisée]
52. Guo, Y. ; Du, C. ; Maman, Z. ; Chen, X. ; Yu, K. VoiceFlow : synthèse vocale efficace avec correspondance de flux rectifiée. Dans les actes de la conférence internationale IEEE 2024 sur l'acoustique, la parole et le traitement du signal (ICASSP2024), Séoul, République de Corée, 14-19 avril 2024 ; pp. [Référence croisée]

53. Wang, C. ; Chen, S. ; Wu, Y. ; Zhang, Z. ; Zhou, L. ; Liu, S. ; Chen, Z. ; Liu, Y. ; Wang, H. ; Li, J. ; et coll. Modèles de langage de codec neuronal sont des synthétiseurs de synthèse vocale Zero-Shot. arXiv 2023, arXiv :2301.02111.
54. Lajszczak, M. ; Cámbara, G. ; Li, Y. ; Beyhan, F. ; van Korlaar, A. ; Yang, F. ; Joly, A. ; Martín-Cortinas, Á.; Abbas, A. ; Michalski, A. ; et coll. BASE TTS : leçons tirées de la création d'un modèle de synthèse vocale comportant un milliard de paramètres sur 100 000 heures de données. arXiv 2024, arXiv :2402.08093.