Ciencias Aplicadas

MDPI

Artículo

Uso del aprendizaje por transferencia para realizar Dungan de bajos recursos Síntesis del habla del lenguaje

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

[1] Facultad de Ingeniería Electrónica y de la Información, Universidad de Tongji, Shanghai 201804, China; liuxh709@163.com

[2] Escuela de Tecnología Educativa, Universidad Normal del Noroeste, Lanzhou 730070, China; jiangh940618@163.com

[3] Laboratorio clave de digitalización de la educación de la provincia de Gansu, Lanzhou 730070, China

* Correspondencia: yanghw@nwnu.edu.cn

† Estos autores contribuyeron igualmente a este trabajo.

Resumen: Este artículo presenta un método basado en el aprendizaje por transferencia para mejorar la calidad del habla sintetizada del idioma Dungan de bajos recursos. Esta mejora se logra ajustando un modelo acústico en mandarín previamente entrenado a un modelo acústico en lenguaje Dungan utilizando un corpus Dun-gan limitado dentro del marco Tacotron2+WaveRNN. Nuestro método comienza con el desarrollo de un analizador de texto Dungan basado en transformador capaz de generar secuencias unitarias con información prosódica incorporada a partir de oraciones Dungan. Estas secuencias de unidades, junto con las características del habla, proporcionan pares de <secuencia de unidades con etiquetas prosódicas, espectrogramas de Mel> como entrada de Tacotron2 para entrenar el modelo acústico. Al mismo tiempo, entrenamos previamente un modelo acústico mandarín basado en Tacotron2 utilizando un corpus mandarín a gran escala. Luego, el modelo se afina con un corpus de voz de Dungan a pequeña escala para derivar un modelo acústico de Dungan que aprende de forma autónoma la alineación y el mapeo de las unidades con los espectrogramas. Los espectrogramas resultantes se convierten en formas de onda a través del codificador de voz WaveRNN, lo que facilita la síntesis de habla mandarín o dungan de alta calidad. Tanto los experimentos subjetivos como los objetivos sugieren que la síntesis de voz de Dungan basada en el aprendizaje por transferencia propuesta logra puntuaciones superiores en comparación con los modelos entrenados sólo con el corpus de Dungan y otros métodos. En consecuencia, nuestro método ofrece una estrategia para lograr la síntesis del habla en lenguas de bajos recursos agregando información prosódica y aprovechando un corpus lingüístico similar de altos recursos a través del aprendizaje por transferencia.

Palabras clave: Síntesis del habla en lengua Dungan; análisis de texto; transferir aprendizaje; lenguaje de bajos recursos ; tacotrón2

Cita: Liu, M.; Jiang, R.; Yang, H. Uso del aprendizaje por transferencia para realizar la síntesis del habla en lengua dungan de bajos recursos. Aplica. Ciencia. 2024, 14, 6336. https://doi.org/10.3390/app14146336

Editores Académicos: Gloria Corpas Pastor y Tharindu Ranasinghe

Recibido: 17 de junio de 2024
Revisado: 17 de julio de 2024
Aceptado: 18 de julio de 2024
Publicado: 20 de julio de 2024

1. Introducción

La síntesis de voz (conversión de texto a voz (TTS)) se utiliza ampliamente en hogares inteligentes, sistemas de navegación y aplicaciones de audiolibros. A nivel mundial, existen aproximadamente 6.000 idiomas , la mayoría considerados de bajos recursos. Si bien se han logrado avances significativos en la síntesis de voz para los principales idiomas, como el mandarín, el inglés y el francés, la calidad de voz de TTS para los idiomas de bajos recursos, como el tibetano y el dungan, sigue siendo subóptima. En los últimos años, ha habido un aumento en la investigación centrada en la síntesis del habla en lenguajes de bajos recursos, como lo demuestran numerosos estudios [1–6]. Sin embargo, aún es necesario completar la investigación sobre la síntesis del habla en lengua dungan . La lengua dungan, que es una variante de los dialectos Shanxi-Gansu dentro del dialecto chino hablado en Asia Central, está clasificada como una lengua de bajos recursos debido a su uso limitado, el número cada vez menor de hablantes y la escasez de materiales lingüísticos [7, 8]. Dado que el ruso se ha convertido en el idioma oficial de Asia Central, la creación de un corpus de habla integral con conocimientos lingüísticos para una síntesis de voz Dungan de alta calidad presenta un desafío importante. Aun

En la síntesis de voz de Dungan basada en DNN [9,10], la calidad del habla sintetizada no fue alta debido al corpus de entrenamiento limitado.

Las tecnologías de síntesis de voz abarcan la síntesis de voz concatenativa basada en la selección de unidades [11], la síntesis de voz paramétrica estadística (SPSS) basada en el modelo oculto de Markov (HMM) [12] y la síntesis de voz basada en el aprendizaje profundo [13,14]. Si bien el aprendizaje profundo ha avanzado significativamente la tecnología de síntesis de voz, métodos como la memoria a corto plazo (LSTM) y la LSTM bidireccional [15,16] han abordado las limitaciones de la información temporal. Además, los modelos de síntesis de voz de extremo a extremo [17] como Tacotron [18] y Tacotron2 [19] han demostrado la capacidad de asignar texto directamente a voz. Cuando se entrenan con pares de texto a voz a gran escala, estos modelos producen voz sintetizada utilizando codificadores de voz de alta calidad como el algoritmo Griffin-Lim [20], WaveNet [21] y WaveRNN [22].
Sin embargo, tales sistemas requieren importantes corpus de formación. Para los lenguajes de bajos recursos, la falta de corpus de entrenamiento dificulta que los modelos de extremo a extremo aprendan la estructura prosódica de las oraciones, lo que resulta en una falta de cambios prosódicos en el habla sintetizada, lo que afecta su naturalidad, lo que plantea desafíos para la síntesis del habla. de lenguas de bajos recurs

El aprendizaje por transferencia entre idiomas [23-25] se ha empleado para mitigar el problema de la insuficiencia de corpus de capacitación para la síntesis del habla en idiomas de bajos recursos. Esta técnica implica entrenar un modelo de lenguaje utilizando una combinación de un corpus grande de un lenguaje de altos recursos y un corpus más pequeño de un lenguaje de bajos recursos, seguido de adaptar este modelo al lenguaje de bajos recursos. El aprendizaje por transferencia en síntesis de voz ha demostrado ser una estrategia eficaz para producir habla en lenguas de bajos recursos aprovechando las capacidades de un modelo acústico de lengua de altos recursos [26,27].

En nuestra investigación anterior sobre la síntesis del habla tibetana [28-32], determinamos que la integración de información prosódica a través de técnicas basadas en el aprendizaje por transferencia mejora la calidad del habla sintetizada para idiomas de bajos recursos como el tibetano. Sobre la base de esta información, el presente estudio implementa un enfoque de secuencia a secuencia (seq2seq) para la síntesis del habla en lenguaje Dungan, aprovechando el aprendizaje por transferencia y la información prosódica dentro del marco Tacotron2+WaveRNN. Este método implica utilizar un analizador de texto Dungan para extraer etiquetas prosódicas de oraciones Dungan para la integración del modelo, emplear un modelo acústico mandarín basado en Tacotron2 y ajustar el modelo acústico del lenguaje Dungan con un corpus de habla Dungan limitado. Las principales contribuciones se detallan a continuación:

- Front-end: Hemos implementado un analizador de texto completo para el idioma Dungan, que abarca módulos para normalización de texto, segmentación de palabras, predicción de límites prosódicos y generación de unidades basadas en tecnología de transformadores. Este analizador puede producir iniciales y finales como unidades de síntesis de voz con etiquetas prosódicas a partir de oraciones Dungan. • Back-end:

Hemos logrado la síntesis de voz en lenguaje Dungan seq2seq adaptando un modelo acústico mandarín previamente entrenado dentro del marco Tacotron2+WaveRNN.
Esto se logró reemplazando la atención sensible a la ubicación de Tacotron2 con atención directa, mejorando la velocidad de convergencia y la estabilidad.

El resto del artículo se organiza de la siguiente manera. Primero presentamos nuestro marco de síntesis de voz Dungan basado en el aprendizaje por transferencia bajo Tacotron2+WaveRNN en la Sección 2. La configuración experimental y los resultados se presentan en la Sección 3, mientras que los resultados se discuten en la Sección 4. Finalmente, una breve conclusión y un esquema para el trabajo futuro. se proporcionan en la Sección 5.

2. Modelos y métodos

El marco propuesto de síntesis de voz Dungan de bajos recursos basada en el aprendizaje por transferencia, que se muestra en la Figura 1, incluye un módulo de extracción de características, un modelo acústico en mandarín previamente entrenado , un módulo de entrenamiento del modelo acústico Dungan basado en el aprendizaje por transferencia y un sintetizador de voz basado en vocoder WaveRNN.

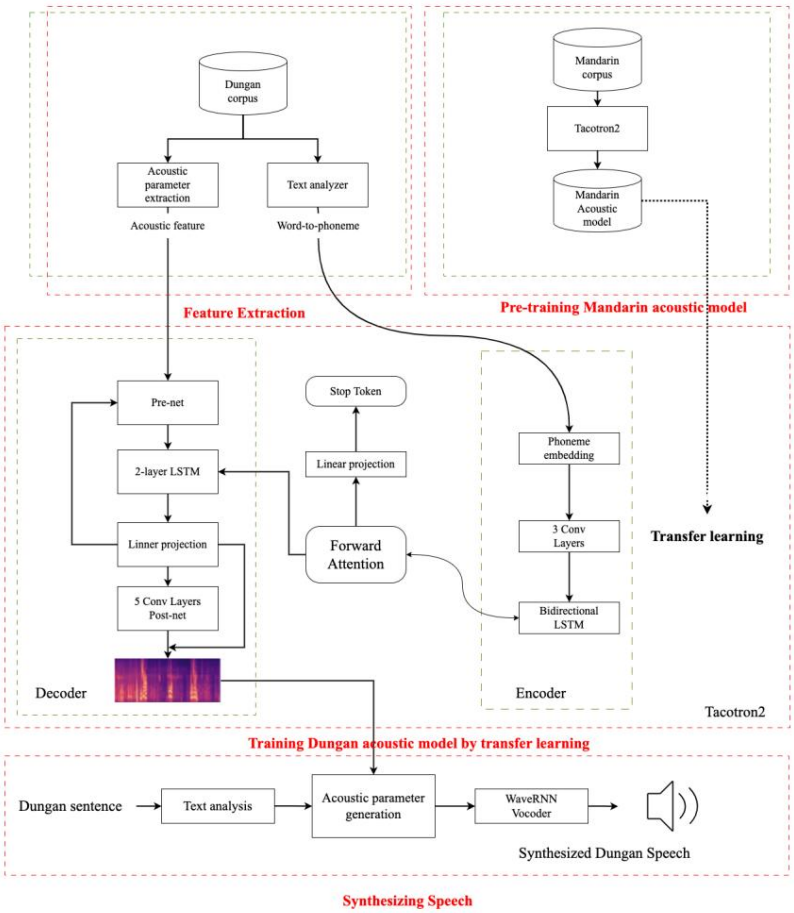

Figura 1. El marco de la síntesis de voz Dungan basada en Tacotron2+WaveRNN.

El módulo de extracción de características extrae características acústicas como el espectrograma Mel de señales de voz y secuencias de unidades de síntesis de voz de oraciones. Hemos desarrollado un analizador de texto completo en lenguaje Dungan para extraer unidades de síntesis de voz con características prosódicas para mapear oraciones Dungan en secuencias de unidades. Dado que tanto el idioma mandarín como el dungan utilizan iniciales y finales como unidades centrales de síntesis del habla, la secuencia de unidades resultante incorpora estos elementos e información prosódica pertinente, incluidos tonos de sílabas y etiquetas de límites prosódicos a nivel de oración.

Dado que Tacotron2 es uno de los marcos de síntesis de voz de codificador a decodificador más populares, y el vocodificador WaveRNN puede generar voz natural, utilizamos Tacotron2 para entrenar modelos acústicos y WAVRNN para convertir espectrograma en forma de onda tanto para el idioma dungan como para el mandarín. El modelo acústico mandarín se entrena previamente con un corpus mandarín a gran escala, mientras que el modelo de lenguaje dungan se transfiere del modelo acústico mandarín con un corpus dungan a pequeña escala.

En la etapa de síntesis de voz, el codificador de voz WaveRNN genera habla dungan o mandarín a partir de la entrada de oraciones dungan o chino. El analizador de texto primero genera las etiquetas dependientes del contexto a partir de la oración de entrada. Luego, las secuencias de la unidad de síntesis de voz (iniciales y finales con su información prosódica) se introducen en el modelo acústico mandarín o dungan para generar el espectrograma Mel. El vocodificador WaveRNN finalmente se utiliza para generar las formas de onda del habla a partir del espectrograma Mel. Utilizamos un analizador de texto chino de fabricación propia para el análisis de texto chino.

2.1. Analizador de texto del idioma Dungan

A diferencia de las técnicas predominantes de síntesis de voz seq2seq diseñadas para los principales idiomas que utilizan únicamente el par <fonema secuencia, habla> para entrenar modelos acústicos, nuestro enfoque emplea una secuencia unitaria que incorpora etiquetas prosódicas como

tono de cada sílaba y el límite prosódico de una oración, sirviendo como la "secuencia de fonemas". En consecuencia, resulta esencial diseñar un analizador de texto integral capaz de extraer las secuencias unitarias de una oración y sus etiquetas prosódicas. Con este fin, aprovechando nuestro analizador de texto chino interno, desarrollamos un analizador de texto en idioma Dungan, como se ilustra en la Figura 2. El proceso comienza con la normalización y segmentación de la oración Dungan de entrada para determinar el límite de las palabras. A esto le sigue un análisis de límites prosódicos para identificar tanto el límite de la palabra prosódica como el de la frase prosódica. En la etapa final, las iniciales y finales de los personajes de Dungan se derivan a través de un proceso de conversión de personajes a unidades basado en transformadores.

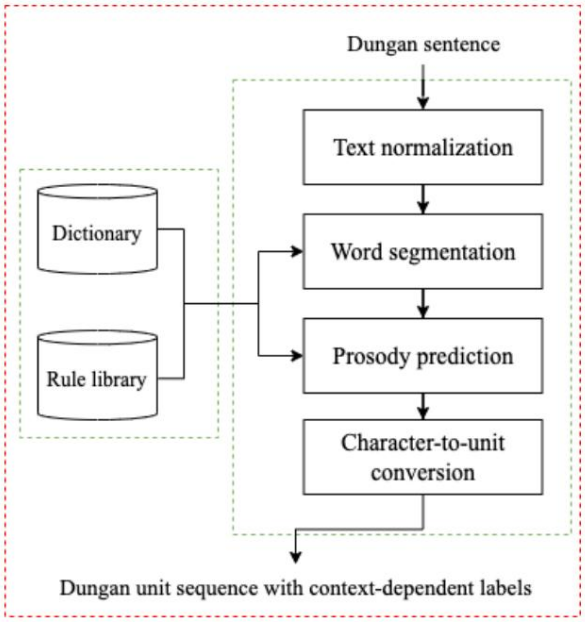

Figura 2. Procedimiento de análisis de texto de Dungan.

2.1.1. Unidad de síntesis del habla de la lengua dungan

A pesar de utilizar un sistema de escritura diferente, Dungan representa una pronunciación dialectal del mandarín fuera de China. El idioma dungan está escrito en escritura cirílica, asemejándose a los idiomas eslavos como el ruso, por lo que el idioma dungan consta de caracteres fonéticos con ortografía secuencial, siguiendo una estructura similar al chino [33-35]. El orden ortográfico de los caracteres Dungan consta de iniciales, finales y tono, como se muestra en la Figura 3.

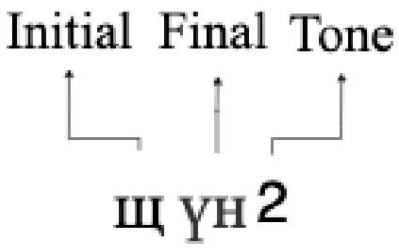

Figura 3. Estructura de un personaje Dungan.

Este artículo utiliza iniciales y finales como unidad de síntesis de voz. El personaje Dungan consta de 25 iniciales (incluida la inicial cero) y 32 finales, como se muestra en la Tabla 1.
Al igual que el mandarín, los tonos del idioma dungan son cruciales para distinguir la semántica y las emociones [36]. Dungan presenta cuatro tonos, excluyendo el tono claro, a saber, el tono de nivel (21), el tono ascendente (24), el tono descendente-ascendente (53) y el tono descendente (44), cada uno indicado por los números del 1 al 4, respectivamente.

Tabla 1. Las iniciales y finales de Dungan Language.

iniciales	/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/, /en/, /yN/

2.1.2. Normalización de texto

Cualquier oración de entrada puede contener formas numéricas de hora, fecha, abreviaturas y sustantivos propietarios especiales. Antes de convertir una oración en una secuencia de símbolos fonéticos, es esencial utilizar la normalización de texto para transformar texto no estándar en un símbolo fonético unificado. Por lo tanto, implementamos una normalización de texto basada en reglas para identificar caracteres que no sean Dungan. Desarrollamos un conjunto de reglas de normalización de texto de Dungan basadas en reglas de normalización de texto chino [37] y empleamos el método agregar-restaurar para normalizar los caracteres de Dungan de acuerdo con [38].

2.1.3. Segmentación de

palabras Los límites de las palabras juegan un papel importante en la predicción de límites prosódicos. Por tanto, es esencial identificar los límites de las palabras de una oración después de la normalización. Las oraciones de Dungan exhiben distinciones claras entre palabras y sílabas, lo que hace que la segmentación sea relativamente sencilla. Empleamos un algoritmo de segmentación de palabras basado en la máxima coincidencia para extraer palabras de Dungan de la oración de entrada. Compilamos un diccionario de palabras de Dungan que comprende 49.293 palabras para facilitar este proceso. La palabra más larga tiene ocho caracteres en este diccionario, mientras que la más corta tiene un solo carácter. El diccionario abarca principalmente términos básicos de Dungan, como se menciona en fuentes como “Diccionario común de la lengua Dungan” [39], “Una encuesta sobre la lengua tungan en Asia central” [40], “Una encuesta sobre la lengua Dungan” [41] y Términos adicionales de búsqueda de Dungan

2.1.4. Predicción de límites prosódicos

Nuestro enfoque utiliza iniciales y finales, junto con sus etiquetas prosódicas, como secuencia de unidades de entrada para el modelo acústico. Por tanto, extraer la estructura prosódica de las oraciones Dun-gan es crucial para sintetizar un discurso de alta calidad. Al igual que el mandarín, la jerarquía prosódica de Dungan se puede segmentar en palabras prosódicas, frases prosódicas, frases de entonación y pausas en las oraciones. Los límites de las frases de entonación se pueden identificar fácilmente utilizando los signos de puntuación de Dungan. En este estudio, empleamos un BiLSTM con un método basado en campo aleatorio condicional (BiLSTM_CRF), como se ilustra en la Figura 4, para predecir los límites de palabras y frases prosódicas [42].

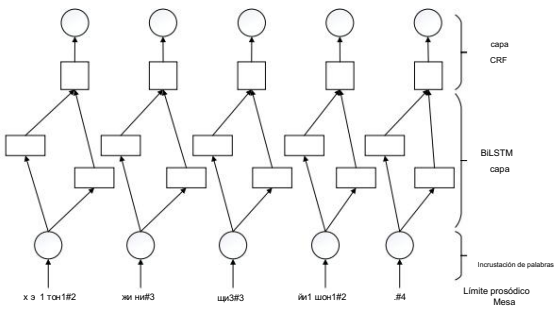

Figura 4. El marco de la predicción de límites prosódicos de Dungan basada en BLSTM_CRF. La entrada es una oración de Dungan con información prosódica.

Empleamos cuatro conjuntos distintos de etiquetado de posición de palabras prosódicas (#1, #2, #3, #4) para categorizar las palabras de Dungan en frases prosódicas. Específicamente, el número 1 se utilizó para denotar las palabras prosódicas, el número 2 designó las frases prosódicas, el número 3 marcó la terminación de un Dungan.

palabra, y el n.° 4 indicó una pausa dentro de una oración. El proceso de etiquetado incorporó frases e información prosódica derivada del texto de Dungan, que se etiquetó manualmente. Durante esta fase, los lingüistas revisaron y modificaron esporádicamente oraciones seleccionadas. Logramos un alto nivel de coherencia con los expertos lingüísticos mediante correcciones iterativas.

A pesar de la capacidad de BiLSTM para aprender información dependiente del contexto, sus decisiones de clasificación independientes están limitadas por fuertes dependencias en la etiqueta de salida. Para abordar esto, empleamos una capa CRF que considera etiquetas vecinas, como se ilustra en la Figura 4. Para una oración de entrada normalizada X = {x1, x2, · · · , xn} que contiene n palabras y una secuencia de etiquetas de oración y = (y1, y2, . , yn), cada palabra se representa como un vector d-dimensional mediante word2vec. Definimos su puntuación de predicción s(X, y) de la siguiente manera:

$$(X, y) = \sum_{yo=1}^{norte} Pi\text{-}yi + \sum_{yo=0}^{norte} Ay\ ,yi+1 \quad (1)$$

donde P es la matriz de puntuaciones generadas por la red BLSTM. Pi,yi corresponde a la puntuación de la etiqueta yi de la iésima palabra de una oración. A es la matriz de puntuaciones de transición de la capa CRF, y Ayi ,yi+1 corresponde a la puntuación de la etiqueta yi a la etiqueta yi+1 .

En el entrenamiento, maximizamos las siguientes funciones de probabilidad logarítmica:

$$\text{Iniciar sesión}(p(y \mid X)) = s(X, y) - \text{Iniciar sesión} \sum_{y \quad YX}^{} {}_{mi}^{s(X,y)} \quad (2)$$

donde YX representa todas las secuencias de etiquetas posibles para un texto de entrada X. se

En la decodificación, la secuencia óptima y proporciona de la siguiente

$$y \quad s(X, y) = \underset{y \quad YX}{\operatorname{argmax}} \quad \text{manera:} \quad (3)$$

2.1.5. Conversión de carácter a unidad basada en transformador

Mandarín y Dungan emplean el mismo sistema Pinyin para etiquetar la pronunciación. En consecuencia, la conversión de carácter a unidad en Dungan es paralela a la del mandarín. Este estudio introduce un enfoque basado en transformadores [43] para derivar la unidad Dungan, como se ilustra en la Figura 5, para mejorar la precisión de la conversión de carácter a unidad de Dungan. El codificador y el decodificador se forman apilando las mismas capas esenciales con N = 6. Cada capa subyacente consta de dos subcapas. La primera subcapa es la capa de atención de múltiples cabezas. El decodificador tiene una capa de atención multicabezal oculta (atención multicabezal enmascarada).

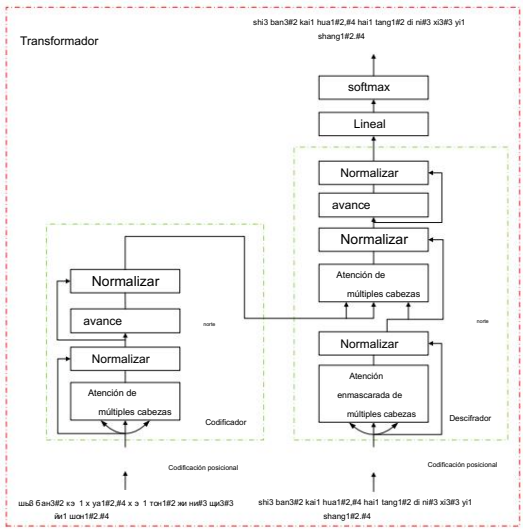

Figura 5. El marco de la conversión de carácter a unidad de Dungan basada en Transformer. La entrada es una oración Dungan con información prosódica (izquierda) y su correspondiente secuencia Pinyin (derecha). El resultado es la secuencia Pinyin con información prosódica.

2.2. Transferir el modelo acústico de Dungan basado en el

aprendizaje Implementamos el modelo acústico de Dungan ajustando un mandarín previamente entrenado modelo acústico, como se ilustra en la Figura 6.

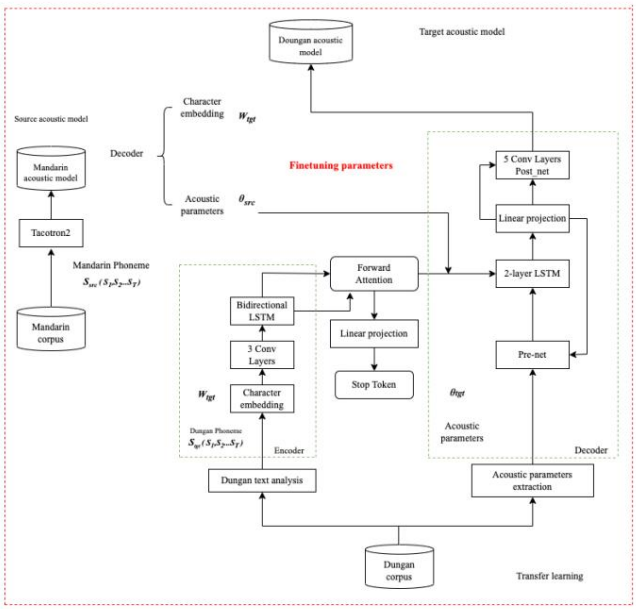

Figura 6. Procedimiento de entrenamiento del modelo acústico del lenguaje Dungan con aprendizaje por transferencia.

2.3. Modelo acústico mandarín basado en Tacotron2 previamente entrenado

El modelo acústico mandarín se entrena inicialmente utilizando un corpus mandarín a gran escala. Nuestro analizador de texto chino patentado extrae las iniciales, finales y etiquetas prosódicas asociadas de estas oraciones. Las características acústicas extraídas abarcan el espectrograma Mel del corpus mandarín a gran escala dentro del marco Tacotron2.

Dada la pronunciación similar entre el idioma Dungan y el mandarín, empleamos el método de aprendizaje de transferencia de mapeo [44] para obtener un modelo acústico Dungan (idioma de destino) mediante la transferencia de conocimiento del mandarín (idioma de origen), que se puede formular de la siguiente manera:

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

donde θ son los parámetros del modelo acústico, W denota incrustaciones de símbolos que se pueden aprender e Y representa el espacio del mandarín. XL es el espacio de texto para el idioma Dungan.

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

donde L es la unidad establecida para el idioma Dungan, St es la t-ésima unidad de la secuencia de unidades de Dungan y T es la longitud de la secuencia de unidades.

En el codificador, ingresamos una secuencia de unidades Dungan representada por incrustaciones de caracteres. Esto pasa a través de una pila de tres capas convolucionales, seguido de la normalización por lotes y activaciones ReLU. Posteriormente, la salida de la capa convolucional final se introduce en una capa LSTM bidireccional para generar las características de la unidad Dungan.

El aprendizaje por transferencia basado en mapeo implica mapear instancias de θsrc y θtgt en un nuevo espacio de parámetros acústicos. En este proceso, podemos utilizar directamente Wsrc y θsrc decodificados del modelo acústico mandarín por el decodificador. θsrc y θtgt pueden tomar incrustaciones como entrada y generar voz. Sin embargo, debido a que ssrc y stgt provienen de conjuntos de símbolos diferentes, es decir, Lsrc = Ltgt, el mismo concepto no se puede aplicar directamente a Wsrc y Wtgt. Para abordar este problema, las unidades Dunggan están integradas en Wtgt para facilitar el reaprendizaje durante el proceso de transmisión.

Adoptamos el mecanismo de atención directa, que utiliza pesos de atención acumulativos. para calcular el vector de contexto.

El decodificador es una red neuronal recurrente autorregresiva que predice un θtgt a partir de la secuencia de la unidad Dungan de entrada del codificador un cuadro a la vez. Podemos usar θsrc aprendido del modelo acústico mandarín para inicializar θtgt en el nuevo espacio de parámetros acústicos. La salida del paso de tiempo inicial se procesa primero a través de una red previa que consta de dos capas completamente conectadas. Este resultado se combina con el contexto de atención directa. vector y pasó a través de un par de capas LSTM. La combinación de las salidas LSTM. y los vectores de contexto de atención experimentan tres transformaciones lineales distintas para predecir el marco del espectrograma objetivo, token de parada y residuo estimado. Posteriormente, lo previsto Las características acústicas están sujetas a cinco capas convolucionales, generando un residual para mejorar. la reconstrucción del modelo acústico de Dungan.

3. Resultados

3.1. Evaluación de la conversión de carácter a unidad de Dungan con base de transformador

El análisis de texto en la parte frontal afecta la calidad de la síntesis de voz en la parte posterior. Al final, evaluamos el analizador de texto Dungan, donde la conversión de carácter a unidad El módulo es el factor más crítico que afecta la calidad del habla sintetizada. para evaluar la viabilidad del módulo de conversión de caracteres a unidades de Dungan basado en transformadores, utilizó un conjunto de datos que comprende 10,783 oraciones en el idioma Dungan transcritas usando Mandarín Pinyin. Las representaciones del lenguaje Dungan y del mandarín pinyin del conjunto de datos. son isomórficos y encapsulan atributos textuales como el tono y los límites prosódicos inherentes al idioma dungan. En nuestra investigación asignamos el 10% del total de 10.783 sentencias para servir como conjunto de prueba, otro 10% como conjunto de validación y el 80% restante fueron designado como conjunto de entrenamiento. Los hiperparámetros asociados con el transformador son detallado en la Tabla 2. Empleamos medidas de precisión, recuperación y F1 como índices de evaluación, como ilustrado en la Tabla 3. Los resultados del proceso de evaluación afirmaron que la propuesta El módulo de carácter a unidad de Dungan es adecuado para la evaluación posterior de la síntesis de voz.

Tabla 2. Hiperparámetros del modelo de conversión de carácter a unidad basado en transformador.

Parámetro	Valor
Capas de atención Nx	6
cabezas	8
Tamaño del lote	32
Oculto	513
Abandonar	0.1
Tasa de aprendizaje	0.0001

Tabla 3. Resultados de la conversión de carácter a unidad de Dungan basada en Transformer.

Precisión	Recordar	F1
90.12	89,91	90.01

3.2. Evaluación de modelos acústicos Dungan basados en el aprendizaje por transferencia

3.2.1. Cuerpo

En el experimento, utilizamos grabaciones de nueve mujeres y treinta y un hombres hablantes. de la base de datos china de Tsinghua de 30 horas [45] (con un total de 13.389 frases) como corpus en mandarín. Para el corpus de Dungan, seleccionamos grabaciones de cinco hablantes masculinos (923 por ciento). persona, totalizando 4615 sentencias y 6 h). El corpus de Dungan abarca todos los aspectos iniciales y pronunciaciones finales del idioma Dungan. La longitud promedio de una oración es de 18 sílabas, con una duración promedio de 10 s. Todas las grabaciones se convirtieron a monocanal de 16 kHz. Frecuencia de muestreo con precisión de cuantificación de 16 bits.

3.2.2. Configuración experimental

Tres tipos de marcos TTS, incluidos Tacotron+Griffin-Lim, Tacotron2+WaveNet, y Tacotron2+WaveRNN, se compararon en los experimentos. Algunos hiperparámetros de los marcos se proporcionan en la Tabla 4.

Tabla 4. Hiperparámetros del modelo de Tacotron y Tacotron2.

Modelo		Tacotrón	Tacotrón2	Tacotron2 de atención hacia adelante
codificador de voz		Griffin-Lim	OndaNet	ondaRNN
Codificador	incrustar	Fomema (256)	Fomema (512)	Fomema (512)
	Pre-red	FFN (256, 128)	-	Fomema FFN (512, 256)
	Núcleo del codificador	CBHG (256)	CNN (512) Bi-LSTM (512)	CNN (256) Bi-LSTM (256, 512)
Descifrador	post-red	CBHG (256)	CNN (512)	CNN (512)
	Decodificador RNN	GRU (256, 256)	-	LSTM (512, 256)
	Atención	Aditivo (256)	Sensible a la ubicación (128)	Adelante (256)
	Atención RNN	GRU (256)	LSTM (1024, 1024)	LSTM (256)
	Pre-red	FFN (256, 128)	FFN (256, 256)	FFN (256, 128)
Parámetro		7,6 × 106	28,9 × 106	23,7 × 106

Los tres marcos comprenden un módulo de análisis de texto frontal, un modelo acústico módulo de formación y un vocoder. El módulo analizador de texto transforma dungan o chino oraciones en una secuencia de unidades representadas en Pinyin, incluidas las iniciales, las finales y sus tonos y etiquetas de límites prosódicas. En el módulo de entrenamiento del modelo acústico, derivamos el registro espectrograma de magnitud de la señal de voz usando ventanas de Hann con 80 ms longitud del fotograma, desplazamiento del fotograma de 12,5 ms y transformada de Fourier de 2048 puntos.

Para el marco Tacotron+Griffin-Lim, los modelos acústicos se entrenan utilizando una salida factor de reducción de capa de r = 3 y el optimizador Adam con una tasa de aprendizaje decreciente. El La tasa de aprendizaje comienza en 0,001 y posteriormente se reduce a 0,0005, 0,0003 y 0,0001. después de 5, 20 y 50 épocas, respectivamente. Se emplea una función de pérdida sencilla para Decodificador seq2seq (espectrograma Mel) y la red de posprocesamiento (espectrograma lineal). El tamaño del lote de entrenamiento se establece en 32, y todas las secuencias se rellenan hasta una longitud máxima de reconstruir los marcos rellenos con ceros. El algoritmo Griffin-Lim se utiliza como codificador de voz. para la conversión de espectro a voz de Mel.

Para el marco basado en Tacotron2+WaveNet, entrenamos los modelos acústicos utilizando el procedimiento estándar de entrenamiento de máxima probabilidad, que implica alimentar la salida correcta en lugar de la salida prevista en el lado del decodificador. Esto se completó con un tamaño de lote de 32. El optimizador Adam se utilizó con los parámetros establecidos de la siguiente manera: β = 0,9, β = 0,999, = 10-6 . La tasa de aprendizaje se inicializó en 10-3 y luego descendió exponencialmente a 10-5 después de 50, 000. Además, aplicamos la regularización L2 con un peso de 10−6 . para el mel Para la conversión de espectro a voz, se empleó WaveNet como codificador de voz.

En nuestro marco de aprendizaje por transferencia basado en Tacotron2+WaveRNN, inicialmente empleamos un corpus mandarín a gran escala para entrenar previamente un modelo acústico mandarín para el modelo posterior transferir. Este modelo previamente entrenado se utiliza luego para entrenar el modelo acústico de Dungan mediante transferencia. aprendiendo del corpus mandarín-dungan. Para la codificación de voz, utilizamos WaveRNN para Conversión de espectro a voz de Mel. Dado que la configuración de los parámetros afecta significativamente Precisión y robustez del modelo, optimizamos estos parámetros a través de entrenamiento iterativo. y actualizaciones.

Cada marco TTS implementa una síntesis de voz monolingüe para mandarín o dungan y una bilingüe basada en el aprendizaje por transferencia. Entrenamos varios modelos en tres

Marcos TTS para evaluar la calidad y claridad del discurso sintetizado. En nuestro experimento, El 10% de las expresiones se asignaron aleatoriamente al conjunto de prueba, otro 10% se designó para el conjunto de desarrollo, y las expresiones restantes constituyeron el conjunto de entrenamiento.

Modelo monolingüe dependiente del hablante de Dungan

Entrenamos el modelo acústico Dungan Monolingual Speaker-Dependent (DSD) utilizando grabaciones de cinco hablantes masculinos, cada uno de los cuales contribuyó con 923 oraciones, con un total de 4615. frases y una duración de 6 h. Luego comparamos la calidad y claridad de los sintetizados. discurso en tres marcos: DSD-Tacotron+Griffin-Lim, DSD Tacotron2+WaveNet, y DSD-Tacotron2+WaveRNN.

Modelo dependiente del hablante monolingüe en mandarín

Utilizamos grabaciones de nueve hablantes mujeres y treinta y un hombres (Tsinghua Base de datos china de 30 horas, compuesta por 13.389 frases) para entrenar el mandarín monolingüe Modelo acústico dependiente del altavoz (MSD). Comparamos la calidad del habla sintetizada. y claridad en tres marcos: MSD-Tacotron+Griffin-Lim, MSD-Tacotron2+WaveNet, y MSD-Tacotron2+WaveRNN.

Modelo dependiente del hablante bilingüe mandarín y dungan

Utilizamos grabaciones de cinco hablantes masculinos de Dungan (923 oraciones por individuo, sumando hasta 4615 oraciones, equivalente a 6 h) como datos de entrenamiento para transferir el modelo acústico mandarín al modelo acústico Dungan para realizar un altavoz dependiente de Dungan (MDSD) y un modelo acústico dependiente del hablante mandarín (MDSM). Nosotros Luego compararon la calidad y claridad del habla sintetizada en seis marcos.

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

3.2.3. Evaluaciones objetivas

Empleamos la distorsión Mel-cepstral (MCD) [46], distorsión de periodicidad de banda A (BAP) [47], error cuadrático medio (RMSE) [48] y error sonoro/sordo (V/UV) [47] evaluar objetivamente los distintos modelos. Los resultados de las pruebas acústicas DSD y MSD. Los modelos se presentan en la Tabla 5 y la Tabla 6, respectivamente. De manera similar, el MDSM y el MDSD Los resultados se muestran en la Tabla 7 y la Tabla 8, respectivamente.

Tabla 5. Resultados objetivos del modelo acústico DSD para Dungan.

Modelo	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	9.675	9.572	9.502
PAB (dB)	0,189	0,187	0.170
F0 RMSE (Hz)	32.785	32.692	32.087
V/UV (%)	9.867	9.721	9.875

Tabla 6. Resultados objetivos del modelo acústico MSD para mandarín.

Modelo	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	5,460	5,291	5.036
PAB (dB)	0,174	0,171	0,169
F0 RMSE (Hz)	14,629	13,986	13.647
V/UV (%)	5,619	5,793	5.762

Tabla 7. Resultados objetivos del modelo acústico MDSD para Dungan.

Modelo	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD (dB)	7,523	7,419	7.395
PAB (dB)	0,178	0,175	0,174
F0 RMSE (Hz)	26,891	26,753	26.617
V/UV (%)	7,774	7,693	7.607

Tabla 8. Resultados objetivos del modelo acústico MDSM para mandarín.

Modelo	Tacotron+Griffin-Lim	Tacotron2+WaveNet	Tacotron2+WaveRNN
MCD(dB)	5.339	5.241	5.108
PAB (dB)	0,174	0,173	0.171
F0 RMSE (Hz)	13.775	13.326	13.092
V/UV (%)	5.542	5.472	5.481

En el contexto de la síntesis de voz Dungan de bajos recursos, la calidad de la alineación de la atención entre el codificador y el decodificador influye significativamente en la calidad de la síntesis de voz. discurso. Las desalineaciones son evidentes principalmente en la legibilidad, los saltos y la repetición. En consecuencia , empleamos la tasa de enfoque diagonal (DFR) y la tasa de inteligibilidad a nivel de palabra. (IR) [49] para evaluar la legibilidad en idiomas de bajos recursos, como se ilustra en la Tabla 9. El DFR Representa el mapa de atención entre el codificador y el decodificador, sirviendo como un mapa arquitectónico. métrico. El IR mide el porcentaje de palabras de prueba pronunciadas correcta y claramente por humanos, una métrica estándar para evaluar la calidad de la generación de voz de bajos recursos.

Tabla 9. Legibilidad del discurso Dungan sintetizado.

Modelo	RI (%)	RFD (%)
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. Evaluación subjetiva

Para las evaluaciones subjetivas, se seleccionaron aleatoriamente 30 frases del conjunto de prueba. Realizamos tres pruebas: puntuación de opinión media (MOS), puntuación de opinión media de degradación (DMOS) y preferencia AB para evaluar la calidad del habla sintetizada. Nosotros reclutamos 20 hablantes nativos de mandarín y 10 estudiantes internacionales nativos de Dungan (que entendían chinos) como participantes. Estos participantes recibieron capacitación antes de la evaluación formal. Los participantes en mandarín evaluaron los modelos acústicos en mandarín de MSD y MDSM, mientras que los participantes de Dungan evaluaron los modelos acústicos de Dungan de DSD y MDSD. Durante la prueba MOS, los participantes calificaron la naturalidad del habla sintetizada en una escala de 5 puntos. escala. Se presentan las puntuaciones promedio de MOS para el habla sintetizada de Dungan y Mandarín. en las Figuras 7 y 8.

En la prueba DMOS, la expresión sintetizada de cada modelo y el original correspondiente La grabación constaba de un par de archivos de voz. Estas parejas se jugaron aleatoriamente al temas, con el discurso sintetizado precediendo al original. Los participantes tuvieron la tarea con comparar meticulosamente los dos archivos y calificar la similitud del sintetizado discurso al original en una escala de 5 puntos. Una puntuación de 5 indicó que el sintetizado el discurso fue similar al original, mientras que una puntuación de 1 significó una disparidad significativa. Las figuras 9 y 10 muestran las puntuaciones DMOS promedio para Dungan y Mandarín sintetizados. discurso, respectivamente.

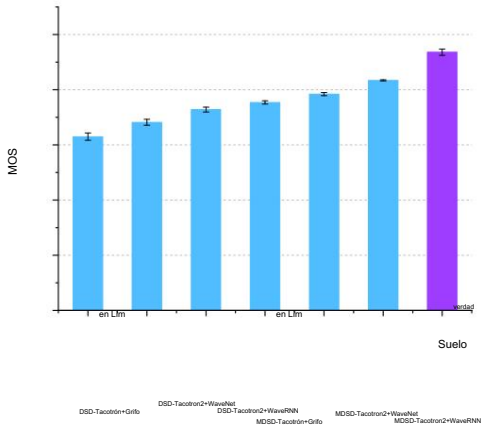

Figura 7. Puntajes MOS promedio del discurso Dungan sintetizado bajo intervalos de confianza del 95%.

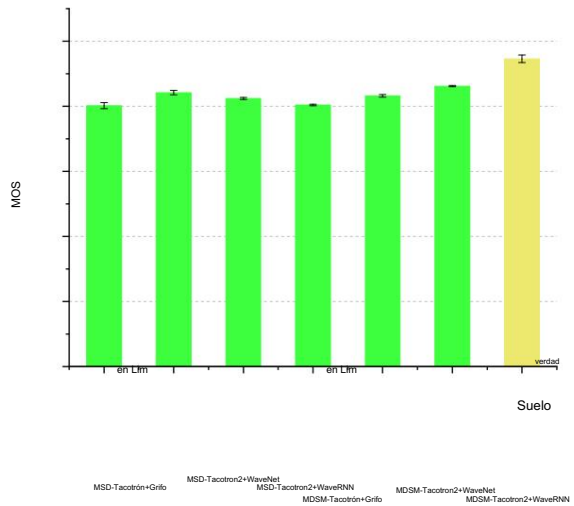

Figura 8. Puntajes promedio de MOS del habla mandarín sintetizada bajo intervalos de confianza del 95%.

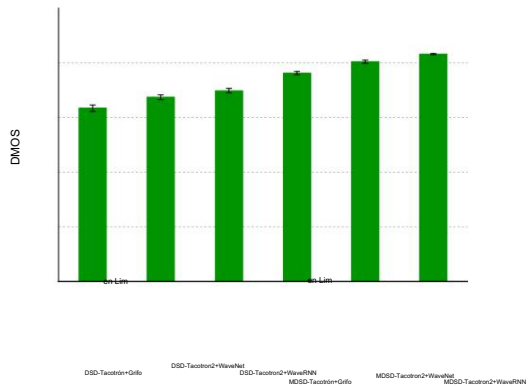

Figura 9. Puntajes DMOS promedio del habla Dungan sintetizada bajo intervalos de confianza del 95%.

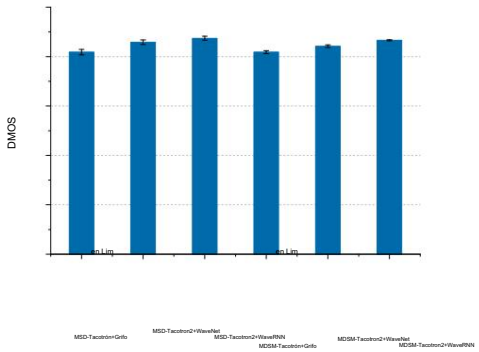

Figura 10. Puntajes DMOS promedio del habla mandarín sintetizada bajo intervalos de confianza del 95%.

En la prueba de preferencia AB, cada par constaba de dos frases idénticas. Las expresiones sintetizadas se reprodujeron en orden aleatorio. Se pidió a los participantes que escucharan y evaluar qué expresión tuvo una calidad superior o indicar “neutral” si no hay preferencia fue discernido. Los resultados sintetizados de preferencia de habla dungan y mandarín son presentados en las Tablas 10 y 11, respectivamente.

Tabla 10. Puntuación subjetiva de preferencia AB (%) de Dungan con ρ < 0,01.

	DSD-Tacotrón+ Griffin-Lim	DSD-Tacotrón2+ OndaNet	DSD-Tacotron2 +WaveRNN	MDSD-Tacotrón+ Griffin-Lim	MDSD-Tacotron2 +WaveNet	MDSD-Tacotrón2+ ondaRNN	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

Tabla 11. Puntuación subjetiva de preferencia AB (%) de mandarín con ρ < 0,01.

	MSD-Tacotrón+ Griffin-Lim	MSD-Tacotrón2+ OndaNet	MSD-Tacotrón2+ ondaRNN	MDSM-Tacotrón+ Griffin-Lim	MDSM-Tacotrón2+ OndaNet	MDSM-Tacotrón2+ ondaRNN	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. Discusión

En evaluaciones objetivas, aunque el marco TTS basado en Tacotron+Griffin-Lim Si se asignan características lingüísticas a características acústicas cuadro por cuadro a través del corpus monolingüe de Dungan, el habla Dungan sintetizada necesita mejorar su calidad y legibilidad. Sin embargo, la atención directa y el modelo acústico afinado pueden mejorar la legibilidad. y reducir el tiempo de entrenamiento. En consecuencia, el Tacotron2+WaveRNN basado en el aprendizaje por transferencia El modelo acústico del marco supera a los demás. Los resultados objetivos del modelo acústico MDSD superan a los del modelo acústico DSD. Esto se debe a que Dungan es una variación. del dialecto del noroeste de China, que comparte muchas similitudes internas. Dadas las similitudes de pronunciación entre mandarín y dungan, el mismo símbolo representa su pronunciaciones exactas. Por lo tanto, concluimos que agregar un corpus mandarín y usar El aprendizaje por transferencia puede mejorar la calidad y legibilidad del habla Dungan sintetizada.

Todas las evaluaciones subjetivas se alinean con evaluaciones objetivas en varios aspectos. El El marco Tacotron2+waveRNN basado en el aprendizaje por transferencia produce una calidad de voz superior,

particularmente en lo que respecta a la naturalidad y legibilidad del habla sintetizada. Con la adición del corpus mandarín, la calidad y legibilidad del habla Dungan sintetizada utilizando los marcos TTS basados en el aprendizaje por transferencia superan a los marcos TTS entrenados en corpus monolingües. Esto se valida aún más mediante la prueba de preferencia AB, que confirma que nuestros marcos TTS propuestos ofrecen calidad y legibilidad mejoradas en comparación con el habla sintetizada por el modelo acústico monolingüe.

5. Conclusiones

Este estudio amplía nuestra investigación anterior mediante la implementación de una síntesis de voz en mandarín basada en el aprendizaje por transferencia y una síntesis de voz en Dungan de bajos recursos bajo el marco Tacotron2+WaveRNN. También desarrollamos un completo analizador de textos Dungan. Los experimentos objetivos y subjetivos revelaron que la síntesis de voz Dungan basada en el aprendizaje por transferencia bajo el marco Tacotron2+WaveRNN superó a los métodos alternativos y al marco de síntesis de voz Dungan monolingüe. Además, el aprendizaje por transferencia no comprometió la calidad del habla ni la legibilidad del habla Dungan sintetizada de bajos recursos. Por lo tanto, nuestro enfoque tiene un potencial significativo para desarrollar sistemas de síntesis de voz para lenguas minoritarias de bajos recursos.

Se han logrado numerosos avances en TTS basados en redes neuronales profundas. Hemos notado que recientemente se han propuesto algunos métodos nuevos de síntesis de voz [50-52] . Motivados por los avances recientes en modelos autorregresivos (AR) que emplean arquitecturas de solo decodificador para la generación de texto, varios estudios, como VALL-E [53] y BASE TTS [54], aplican arquitecturas similares a tareas TTS. Estos estudios demuestran la notable capacidad de las arquitecturas de solo decodificador para producir voz con sonido natural. Estos estudios demuestran la notable capacidad de las arquitecturas de solo decodificador para producir voz con sonido natural. Las investigaciones futuras se centrarán en el uso de estos nuevos métodos para mejorar la calidad de la síntesis de voz en lenguas Dungan, reducir el tamaño del corpus de Dungan y lograr la síntesis de voz para lenguas Dungan utilizando un corpus más grande. Además, se explorará el aprendizaje multitarea para realizar escenarios independientes del hablante y mejorar la emoción del habla Dungan sintetizada.

Contribuciones de los autores: Conceptualización, ML y HY; análisis formal, HY y RJ; curación de datos , ML y RJ; redacción: preparación del borrador original, ML y RJ; redacción: revisión y edición, HY y ML; supervisión, HY; adquisición de financiación, HY Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: La investigación cuenta con el apoyo del fondo de investigación de la Fundación Nacional de Ciencias Naturales de China (Subvención No. 62067008).

Declaración de la Junta de Revisión Institucional: No aplicable a estudios que no involucren a humanos o animales.

Declaración de Consentimiento Informado: No aplicable.

Declaración de disponibilidad de datos: utilizamos dos conjuntos de datos de entrenamiento en el manuscrito. Uno es un conjunto de datos en mandarín disponible públicamente (THCHS-30) y el otro es un conjunto de datos Donggan, que incluye voz y texto. El primero ha sido público y se puede acceder a él desde http://www.openslr.org/18/ (consultado el 16 de junio de 2024). Este último es un conjunto de datos elaborado por uno mismo y no está disponible públicamente. Sin embargo, los datos estarán disponibles previa solicitud.

Conflictos de intereses: Los autores declaran no tener ningún conflicto de intereses. Los financiadores no tuvieron ningún papel en el diseño del estudio; en la recopilación, análisis o interpretación de datos; en la redacción del manuscrito; o en la decisión de publicar los resultados.

Referencias

1. Tu, T.; Chen, YJ; Chieh Yeh, C.; Yi Lee, H. Conversión de texto a voz de un extremo a otro para idiomas de bajos recursos mediante transferencia entre idiomas Aprendiendo. arXiv 2019, arXiv:1904.06508.
2. Liu, R.; Sisman, B.; Bao, F.; Yang, J.; Gao, G.; Li, H. Explotación de las características morfológicas y fonológicas para mejorar el fraseo prosódico para la síntesis del habla mongol. Trans. IEEE/ACM. Idioma de voz en audio. Proceso. 2021, 29, 274–285. [Referencia cruzada]
3. Saeki, T.; Maití, S.; Li, X.; Watanabe, S.; Takamichi, S.; Saruwatari, H. Adaptación del lenguaje basado en gráficos inductivos de texto para síntesis del habla de bajos recursos. Trans. IEEE/ACM. Idioma de voz en audio. Proceso. 2024, 32, 1829–1844. [Referencia cruzada]

4. Xu, J.; Bronceado, X.; Ren, Y.; Qin, T.; Li, J.; Zhao, S.; Liu, TY LRSpeech: Síntesis y reconocimiento del habla con recursos extremadamente bajos. En Actas de la 26.ª Conferencia Internacional ACM SIGKDD sobre Descubrimiento de Conocimiento y Minería de Datos, KDD'20, Nueva York, NY, EE. UU., 6 a 10 de julio de 2020; págs. 2802–2812. [Referencia cruzada]
5. Él, M.; Yang, J.; Él, L.; Soong, FK Modelos multilingües Byte2Speech para síntesis de voz escalable y de bajos recursos. arXiv 2021, arXiv:2103.03541.
6. Oliveira, FS; Casanova, E.; Júnior, AC; Soares, AS; Galvão Filho, AR CML-TTS: un conjunto de datos multilingüe para la síntesis del habla en idiomas de bajos recursos. En texto, discurso y diálogo; Ekštein, K., Pártl, F., Konopík, M., Eds.; Springer: Cham, Suiza, 2023; págs. 188-199.
7. Zhu, Y. Idioma Donggan: una variedad especial de los dialectos de Shaanxi y Gansu. Lengua asiática. Culto. 2013, 4, 51–60. 8. Jiang, Y. El idioma Donggan y su relación con los dialectos de Shaanxi y Gansu. J. Chin. Lingüista. 2014, 42, 229–258.
9. Chen, L.; Yang, H.; Wang, H. Investigación sobre la síntesis de voz de Dungan basada en Deep Neural Network. En actas del 11.° Simposio internacional sobre procesamiento del lenguaje hablado chino (ISCSLP) de 2018, Taipei, Taiwán, 26 a 29 de noviembre de 2018; págs. 46–50. [Referencia cruzada]
10. Jiang, R.; Chen, C.; Shan, X.; Yang, H. Uso de la mejora del habla para realizar la síntesis del habla de lenguas dungan de bajos recursos. En Actas de la 24.ª Conferencia de 2021 del Comité Internacional Oriental COCOSDA para la Coordinación y Estandarización de Bases de Datos del Habla y Técnicas de Evaluación (O-COCOSDA), Singapur, 18-20 de noviembre de 2021; págs. 193-198. [Referencia cruzada]
11. Cazar, AJ; Black, AW Selección de unidades en un sistema de síntesis de voz concatenativa que utiliza una gran base de datos de voz. En Actas de la Conferencia Internacional IEEE de 1996 sobre Acústica, Habla y Procesamiento de Señales, Atlanta, GA, EE.UU., 9 de mayo de 1996; Volumen 1, págs. 373–376.
12. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Síntesis de voz basada en modelos ocultos de Markov. Proc. IEEE 2013, 101, 1234–1252. [Referencia cruzada]
13. Ling, ZH; Deng, L.; Yu, D. Modelado de envolventes espectrales utilizando máquinas Boltzmann restringidas y redes de creencias profundas para la síntesis estadística paramétrica del habla. Traducción IEEE. Idioma de voz en audio. Proceso. 2013, 21, 2129–2139. [Referencia cruzada]
14. Zen, H.; Mayor, A.; Schuster, M. Síntesis de voz paramétrica estadística utilizando redes neuronales profundas. En Actas de la Conferencia Internacional IEEE de 2013 sobre Acústica, Habla y Procesamiento de Señales, Vancouver, BC, Canadá, 26 a 31 de mayo de 2013; págs. 7962–7966. [Referencia cruzada]
15. Wang, P.; Qian, Y.; Soong, FK; Él, L.; Zhao, H. Incrustación de palabras para síntesis TTS basada en redes neuronales recurrentes. En Actas de la Conferencia Internacional IEEE sobre Acústica, Habla y Procesamiento de Señales (ICASSP) de 2015, South Brisbane, QLD, Australia, 19 a 24 de abril de 2015; págs. 4879–4883. [Referencia cruzada]
16. Yu, Q.; Liu, P.; Wu, Z.; Ang, SK; Meng, H.; Cai, L. Aprendizaje de información multilingüe con BLSTM multilingüe para la síntesis de voz de idiomas de bajos recursos. En Actas de la Conferencia Internacional IEEE sobre Acústica, Habla y Procesamiento de Señales (ICASSP) de 2016, Shanghai, China, 20 a 25 de marzo de 2016; págs. 5545–5549. [Referencia cruzada]
17. Bronceado, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; Él, L.; et al. NaturalSpeech: síntesis de texto a voz de un extremo a otro con calidad de nivel humano. Traducción IEEE. Patrón Anal. Mach. Intel. 2024, 46, 4234–4245. [Referencia cruzada]
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: hacia la síntesis del habla de un extremo a otro. En Actas de la 18.ª Conferencia Anual de la Asociación Internacional de Comunicación del Habla, Interspeech 2017, Estocolmo, Suecia, 20 a 24 de agosto de 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. Síntesis natural de TTS mediante el acondicionamiento de Wavenet en predicciones de espectrograma MEL. En Actas de la Conferencia Internacional IEEE sobre Acústica, Habla y Procesamiento de Señales (ICASSP) de 2018, Calgary, AB, Canadá, 15 a 20 de abril de 2018; págs. 4779–4783. [Referencia cruzada]
20. Grifo, D.; Lim, J. Estimación de señal a partir de la transformada de Fourier de corto tiempo modificada. Traducción IEEE. Acústico. Proceso de señal de voz. 1984, 32, 236–243. [Referencia cruzada]
21. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Tumbas, A.; Kalchbrenner, N.; Mayor, A.; Kavukcuoglu, K. WaveNet: un modelo generativo para audio sin formato. arXiv 2016, arXiv:1609.03499.
22. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Nouri, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Van den Oord, A.; Dieleman, S.; Kavukcuoglu, K. Síntesis de audio neuronal eficiente. arXiv 2018, arXiv:1802.08435.
23. Byambadorj, Z.; Nishimura, R.; Ayush, A.; Ohta, K.; Kitaoka, N. Sistema de texto a voz para lenguajes de bajos recursos que utiliza aprendizaje por transferencia multilingüe y aumento de datos. EURASIP J. Audio Discurso Música. Proceso. 2021, 2021, 42. [Referencia cruzada]
24. Joshi, R.; Garera, N. Adaptación rápida del hablante en sistemas de texto a voz de bajos recursos utilizando datos sintéticos y aprendizaje por transferencia. En Actas de la 37.ª Conferencia de Asia Pacífico sobre Lenguaje, Información y Computación, Hong Kong, China, 2 a 4 de diciembre de 2023; Huang, CR, Harada, Y., Kim, JB, Chen, S., Hsu, YY, Chersoni, EAP, Zeng, WH, Peng, B., Li, Y., et al., Eds.; LCA: Hong Kong, China, 2023; págs. 267-273.
25. Hacer, P.; Coler, M.; Dijkstra, J.; Klabbers, E. Estrategias en el aprendizaje por transferencia para la síntesis del habla con bajos recursos: mapeo telefónico, entrada de funciones y selección del idioma de origen. En Actas del 12.º Taller de síntesis del habla de ISCA (SSW2023), Grenoble, Francia, 26 a 28 de agosto de 2023; págs. 21-26. [Referencia cruzada]

26. Azizah, K.; Jatmiko, W. Transferencia de aprendizaje, control de estilo y pérdida de reconstrucción del hablante para hablantes múltiples multilingües de disparo cero conversión de texto a voz en idiomas de bajos recursos. Acceso IEEE 2022, 10, 5895–5911. [Referencia cruzada]
27. Cai, Z.; Yang, Y.; Li, M. Síntesis de voz multilingüe y multilingüe con datos limitados de entrenamiento bilingüe. Computadora. Idioma del habla. 2023, 77, 101427. [Referencia cruzada]
28. Yang, H.; Oura, K.; Wang, H.; Gan, Z.; Tokuda, K. Uso del entrenamiento adaptativo del hablante para realizar el intercambio lingüístico mandarín-tibetano síntesis del habla. Multimed. Herramientas Aplica. 2015, 74, 9927–9942. [Referencia cruzada]
29. Wang, L.; Yang, H. Método de segmentación de palabras tibetanas basado en el modelo bilstm_ crf. En actas de la Conferencia internacional IEEE 2018 sobre procesamiento de idiomas asiáticos (IALP), Bandung, Indonesia, 15 a 17 de noviembre de 2018; págs. 297–302.
30. Zhang, W.; Yang, H.; Bu, X.; Wang, L. Aprendizaje profundo para la síntesis de voz translingüe mandarín-tibetano. Acceso IEEE 2019, 7, 167884–167894. [Referencia cruzada]
31. Zhang, W.; Yang, H. Mejora de la síntesis del habla tibetana secuencia a secuencia con información prosódica. Transmisión ACM. Asia de bajos recursos. Lang. inf. Proceso. 2023, 22, 6012. [Referencia cruzada]
32. Zhang, W.; Yang, H. Metaaprendizaje para la síntesis del habla interlingüe mandarín-tibetano. Aplica. Ciencia. 2022, 12, 2185. [Referencia cruzada]
33. Hai, F. Un estudio piloto de palabras prestadas en el idioma dungan de Asia Central. Universidad de Xinjiang. J. 2000, 28, 58–63.
34. Lin, T. Características, situación y tendencias de desarrollo de la lengua tung'gan en Asia central. Contemporáneo. Lingüista. 2016, 18, 234–243.
35. Gladney, DC Alteridad relacional: construcción de identidades dungan (hui), uygur y kazaja en China, Asia central y Turquía. Historia. Antropol. 1996, 9, 445–477. [Referencia cruzada]
36. Miao, DX Modelo de enseñanza bilingüe del pueblo Donggan. J.Res. Educativo. Etnia. Menor. 2008, 19, 111-114.
37. Jia, Y.; Huang, D.; Liu, W.; Dong y.; Yu, S.; Wang, H. Normalización del texto en el sistema de conversión de texto a voz en mandarín. En Actas de la Conferencia Internacional IEEE de 2008 sobre Acústica, Procesamiento de Habla y Señales, Las Vegas, NV, EE. UU., 31 de marzo a 4 de abril de 2008; páginas. 4693–4696. [Referencia cruzada]
38. Wanmezhaxi, N. Investigación sobre varias cuestiones clave en la segmentación de palabras tibetanas. J. Chin. inf. Proceso. 2014, 28, 132-139.
39. Zavyalova, O. Lengua Dungan. 2015. Disponible en línea: https://www.academia.edu/42869092/Dungan_Language (accedido el 16 de junio de 2024).
40. Lin, T. Donggan Writing: una prueba exitosa de escritura alfabética china. J. Segundo. Universidad del Noroeste. Nacional. 2005, 2005, 31–36.
41. Yang, WJ; Zhang, R. Identidad étnica en el contexto transnacional: un caso de estudio de "Dunggan" y la nacionalidad Hui. J. Centro Sur. Univ. Nacional. 2009, 29, 31–36.
42. Zheng, Y.; Tao, J.; Wen, Z.; Li, Y. Predicción de límites prosódicos de extremo a extremo basada en BLSTM-CRF con incrustaciones sensibles al contexto en una interfaz de texto a voz. Proc. Entre discursos 2018, 9, 47–51. [Referencia cruzada]
43. Hlaing, AM; Pa, WP Modelos de secuencia a secuencia para la conversión de grafema a fonema en un gran diccionario de pronunciación de Myanmar. En Actas de la 22.ª Conferencia del Comité Internacional Oriental COCOSDA de 2019 para la Coordinación y Estandarización de Bases de Datos del Habla y Técnicas de Evaluación (O-COCOSDA), Cebú, Filipinas, 25 a 27 de octubre de 2019; págs. 1 a 5. [Referencia cruzada]
44. Tan, C.; Sol, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. Una encuesta sobre aprendizaje por transferencia profunda. En Redes neuronales artificiales y aprendizaje automático: ICANN 2018; K ˚urková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I., Eds.; Springer: Cham, Suiza, 2018; págs. 270–279.
45. Wang, D.; Zhang, X. THCHS-30: Un corpus de libertad de expresión china. arXiv 2015, arXiv:1512.01882.
46. Kubichek, R. Medida de distancia Mel-cepstral para una evaluación objetiva de la calidad del habla. En Actas de la Conferencia IEEE Pacific Rim sobre Computadoras de Comunicaciones y Procesamiento de Señales, Victoria, BC, Canadá, 19 a 21 de mayo de 1993; Volumen 1, págs. 125-128. [Referencia cruzada]
47. Dhiman, JK; Seelamantula, CS Una técnica espectrotemporal para estimar la aperiodicidad y los límites de decisión sonoros/sordos de las señales del habla. En Actas de la Conferencia Internacional IEEE sobre Acústica, Habla y Procesamiento de Señales de 2019 (ICASSP2019), Brighton, Reino Unido, 12 a 17 de mayo de 2019; págs. 6510–6514. [Referencia cruzada]
48. Castelazo, I.; Mitani, Y. Sobre el uso del error cuadrático medio como índice de competencia. Acreditar. Cual. Asegurar. 2012, 17, 95–97. [Referencia cruzada]
49. Ren, Y.; Bronceado, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, TY Texto a voz casi sin supervisión y reconocimiento automático de voz. arXiv 2020, arXiv:1905.06791.
50. Ren, Y.; Hu, C.; Bronceado, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, TY FastSpeech 2: Texto a voz de un extremo a otro, rápido y de alta calidad. arXiv 2022, arXiv:2006.04558.
51. Chen, J.; Canción, X.; Peng, Z.; Zhang, B.; Pan, F.; Wu, Z. LightGrad: Modelo probabilístico de difusión ligera para conversión de texto a voz. En Actas de la Conferencia Internacional IEEE de 2023 sobre Acústica, Habla y Procesamiento de Señales (ICASSP2023), Isla de Rodas, Grecia, 4 a 10 de junio de 2023; págs. 1 a 5. [Referencia cruzada]
52. Guo, Y.; Du, C.; Mamá, Z.; Chen, X.; Yu, K. VoiceFlow: conversión de texto a voz eficiente con coincidencia de flujo rectificada. En Actas de la Conferencia Internacional IEEE de 2024 sobre Acústica, Habla y Procesamiento de Señales (ICASSP2024), Seúl, República de Corea, 14 a 19 de abril de 2024; págs. 11121-11125. [Referencia cruzada]

53. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Modelos de lenguaje de códec neuronal son sintetizadores de texto a voz Zero-Shot. arXiv 2023, arXiv:2301.02111.
54. Łajszczak, M.; Cámbara, G.; Li, Y.; Beyhan, F.; van Korlaar, A.; Yang, F.; Joly, A.; Martín-Cortinas, Á.; Abbas, A.; Michalski, A.; et al. BASE TTS: Lecciones de la construcción de un modelo de texto a voz de mil millones de parámetros en 100.000 horas de datos. arXiv 2024, arXiv:2402.08093.