EVALUACIÓN DEL RENDIMIENTO DE UN ALGORITMO DE VOICE MORPHING SOPORTADO EN TRANSFORMACIONES LINEALES SEGÚN LA NATURALIDAD DE LA VOZ OBTENIDA

Laura Isabel Camacho Morales
Marbell Palechor Alarcón

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telecomunicaciones
Grupo de Nuevas Tecnologías en Telecomunicaciones – GNTT
Procesamiento Digital de Señales
Popayán, 2021

EVALUACIÓN DEL RENDIMIENTO DE UN ALGORITMO DE VOICE MORPHING SOPORTADO EN TRANSFORMACIONES LINEALES SEGÚN LA NATURALIDAD DE LA VOZ OBTENIDA

Trabajo de grado presentado como requisito para obtener el título de Ingeniero en Electrónica y Telecomunicaciones

Laura Isabel Camacho Morales
Marbell Palechor Alarcón

Director: Ing. María Manuela Silva Zambrano

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telecomunicaciones
Grupo de Nuevas Tecnologías en Telecomunicaciones – GNTT
Procesamiento Digital de Señales
Popayán, 2021

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ACRÓNIMOS

ANN	*Artificial Neural Networks*, Redes Neuronales Artificiales.
ANOVA	*Analysis of Variance*, Análisis de la Varianza.
DCT	*Discrete Cosine Transform*, Transformada Discreta del Coseno.
DFT	*Discrete Fourier Transform*, Transformada Discreta de Fourier.
DTW	*Dynamic Time Warping,* Deformación de Tiempo Dinámica.
EM	*Expectation-Maximization,* Algoritmo Esperanza-Maximización.
GMM	*Gaussian Mixture Model*, Modelo de Mezcla Gaussiana.
LPC	*Linear Prediction Coding,* Codificación de Predicción Lineal.
LSF	*Line Spectral Frequencies*, Frecuencias Espectrales de Línea.
MCD	*Mel-cepstral distortion*, Distorsión Mel-cepstral.
MFCC	*Mel Frequency Cepstral Coefficients,* Coeficientes Cepstrales en las Frecuencias de Mel.
ML	*Maximum Likelihood*, Máxima Verosimilitud.
MLPG	*Máximum Likelihood Parameter Generation,* Generación de Parámetros de Máxima Verosimilitud.
MMSE	*Minimum Mean Square Error*, Error Cuadrático Medio Mínimo.
MOS	*Mean Opinion Score,* Puntuación de Opinión Media.
PDF	*Probability Density Function*, Función de Densidad de Probabilidad.
PSOLA	*Pitch Synchronous Overlap and Add*, Fragmentación y Traslape de la Señal Sincronizada en Tono.
RBF	*Radial Basis Functions*, Funciones de Base Radial.
RMSE	*Root Mean Square Error*, Error Cuadrático Medio.
SST	*Speech to Speech Translation,* Traducción de Habla a Habla.
TTS	*Text to Speech Synthesis*, Síntesis de Texto a Habla.

CAPÍTULO 1: LA VOZ

La voz, como principal medio de comunicación oral, juega un papel fundamental en la vida diaria de los seres humanos, ya que además de transmitir mensajes por medio de un idioma, transmite información personal como el estado de ánimo y el idiolecto del hablante [1]. En esencia, la voz es generada por la excitación en las cuerdas vocales que se propaga a través de la faringe y las cavidades nasal y bucal.

En el presente capítulo se presentan los conceptos y definiciones para la comprensión del proceso de la producción de voz, los cuales son relevantes para el este trabajo de grado. Para ello, en la sección 1.1 se explica el funcionamiento del aparato fonador, luego, en la sección 1.2 se presenta la clasificación de los sonidos. En la sección 1.3 se describen los parámetros más importantes para el análisis acústico. Además, se presentan consideraciones para el procesamiento de señales de voz en la sección 1.4. Finalmente, en la sección 1.5 se abordan las consideraciones para definir el ancho de banda para muestrear una señal de voz.

1.1. APARATO FONADOR

Mecánicamente, la voz se produce a través de un proceso físico voluntario del aparato fonador. Éste está compuesto por los pulmones, que son la fuente del flujo de aire; la laringe, donde se encuentran las cuerdas vocales; la faringe, las cavidades oral y nasal, las cuales se comportan como cavidades resonantes y de las cuales dependen las características acústicas de la señal de voz [2].

El aparato fonador se puede desglosar en tres sistemas [3], como se muestra en la Figura 1.1.

1.1.1. Sistema de Generación: Cavidades Infraglóticas

Es en este sistema es donde se inicia el proceso de producción de la voz. Corresponde al sistema respiratorio, el cual está compuesto por los pulmones, los bronquios, el diafragma y la tráquea. Durante el proceso de inspiración, el diafragma, un músculo plano ubicado en la parte inferior de la caja torácica, desciende para dejar que la caja torácica se expanda y entre aire en los pulmones, llegando hasta los bronquios. Cuando se produce la espiración, el diafragma asciende y la cavidad torácica se contrae, esto obliga a los pulmones a expulsar el aire contenido hacia la tráquea, un órgano compuesto por anillos cartilaginosos, y hacia la laringe, a ese flujo de aire se le denomina *flujo glotal* y contiene la energía

necesaria para generar la onda sonora que atraviesa los órganos fonadores superiores [5].

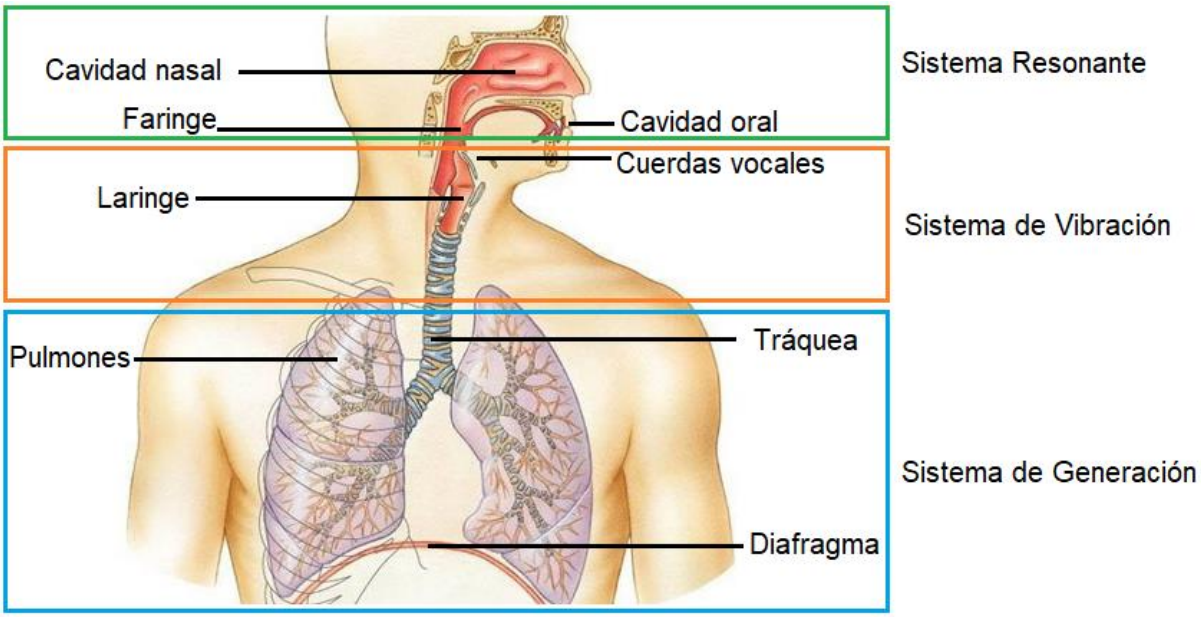


Figura 1.1. Aparato fonador.
Adaptado de [4].

1.1.2. Sistema de Vibración: Cavidad Laríngea

Este sistema es el encargado de transformar el flujo glotal, proveniente de las cavidades infraglóticas, en una señal capaz de excitar las cavidades supraglóticas, es decir, donde se producen los sonidos audibles. Está compuesto principalmente por las cuerdas o pliegues vocales que se descomponen en dos pares, superiores e inferiores. Las cuerdas vocales superiores se denominan también cuerdas vocales vestibulares o falsas ya que no cumplen ninguna función dentro del proceso de generación de la voz, pero sí actúan como protectoras de las cuerdas vocales inferiores y su parte inferior constituye la parte superior del ligamento vestibular que es el que se encarga de evitar que entren objetos extraños como alimentos o saliva en la laringe. Las cuerdas vocales inferiores son las que actúan en la producción de la voz [6]. Durante el proceso de respiración las cuerdas vocales inferiores permanecen separadas para permitir el libre paso de las corrientes de aire, pero, si por el contrario se trata de la producción de sonidos audibles, éstas se unen y tensan haciendo chocar el aire contra ellas.

La laringe se encuentra enseguida y en la parte superior de la tráquea; está conformada por cuatro cartílagos: el cricoides, el último cartílago de la tráquea y la base de la laringe, el tiroides y dos aritenoides. En la parte interna de la laringe se encuentran las cuerdas vocales que están conectadas con el tiroides en su parte inferior y con los cartílagos aritenoides en la parte superior. El movimiento de los

músculos interaritenoideos y cricoaritenoideos provocan que los cartílagos aritenoides cierren y abran las cuerdas vocales. En la Figura 1.2 se muestra una vista longitudinal de la zona donde se encuentran las cuerdas vocales. El espacio libre entre las cuerdas se llama *glotis* [5]. La parte superior de la laringe está unida al hueso hioides.

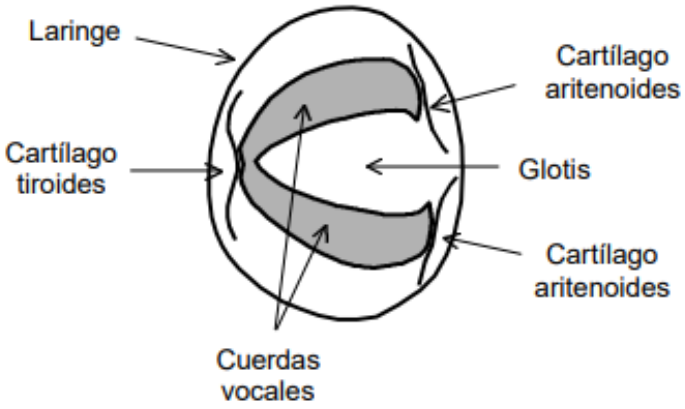


Figura 1.2. Cuerdas vocales.
Adaptado de [2].

1.1.3. Sistema Resonante: Cavidades Supraglóticas

La señal u onda glotal [5], generada en el sistema de vibración llega a las cavidades nasal y oral, donde los órganos articulatorios, mostrados en la Figura 1.3 (labios, lengua, úvula, alvéolo, dientes y fosas nasales), la transforman y amplifican para ser finalmente expulsados al exterior.

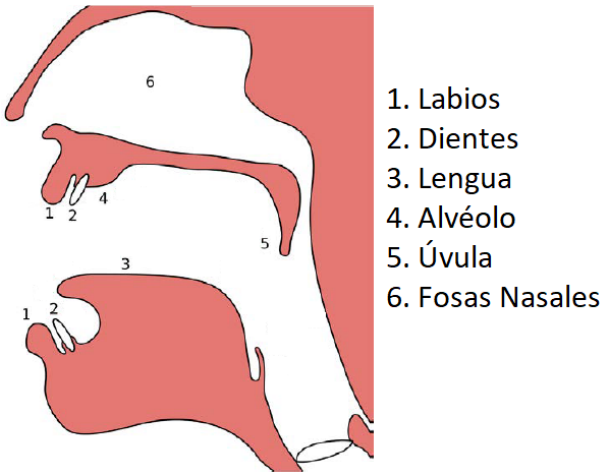


Figura 1. 3. Órganos articulatorios.
Adaptado de [5].

Las cavidades supraglóticas son el conjunto de órganos que participan en este sistema, también son llamadas tracto vocal. Las cavidades supraglóticas están

conformadas por la faringe, la cavidad bucal y la cavidad nasal. En el proceso de producción de la voz tienen la función de perturbar apropiadamente el flujo de aire, proveniente de la cavidad laríngea, para generar la señal acústica expulsada por la nariz y la boca.

La faringe es un conducto tubular que comunica la laringe con las cavidades nasal y bucal, se distinguen en ella tres secciones: nasofaringe, porción nasal de la faringe; bucofaringe, porción oral de la faringe y laringofaringe, porción laríngea de la faringe [7]. Dicho conducto actúa como un resonador y gracias a su estructura *musculoaponeurótica*, le permite movimientos de contracción y relajación, provocando que el aire espirado resuene con mayor o menor intensidad [8]. Los movimientos de la laringe, la lengua y la epiglotis pueden variar la intensidad de la laringofaringe, por el contrario, la intensidad de la bucofaringe obedece a los movimientos de la lengua.

En el tracto vocal existen puntos de articulación (ver Figura 1.4), donde los sonidos cambian en el tiempo, según la posición de los órganos de articulación en el instante dado.

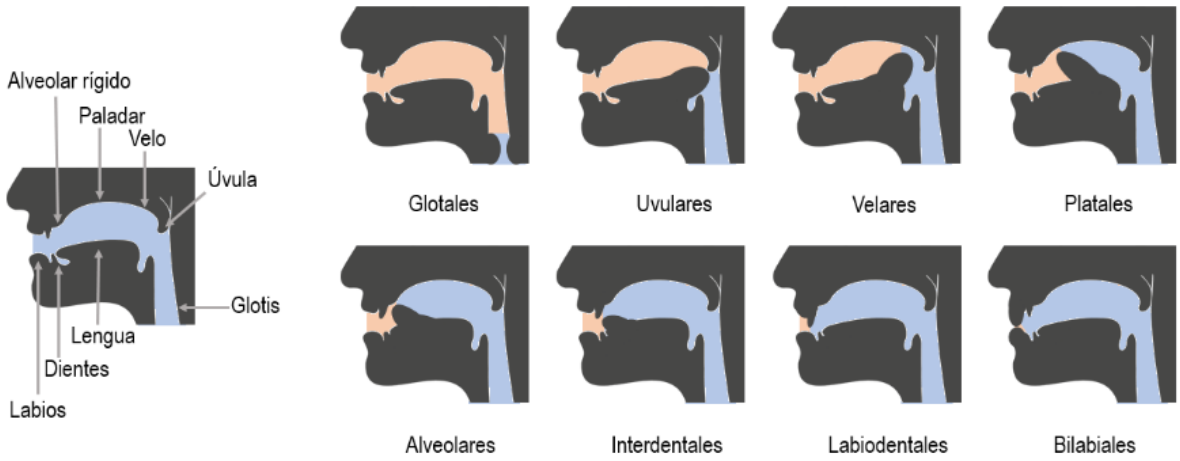


Figura 1.4. Puntos de articulación.
Adaptado de [9].

El proceso de producción de voz comienza cuando el aire sale expulsado de los pulmones en dirección a la laringe. Cuando el aire espirado por los pulmones (flujo glotal) llega a la laringe, donde las cuerdas vocales se encuentran unidas, se produce un aumento en la presión que las obliga a separarse y vibrar. Una vez que se ha liberado la presión, las cuerdas vocales vuelven a unirse, primero la parte inferior y finalmente la parte superior, como se observa en la Figura 1.5. Este fenómeno es conocido como Efecto Bernoulli y se lleva a cabo de manera rápida y repetitiva, por esta razón se le denomina “Ciclo Fonatorio”.

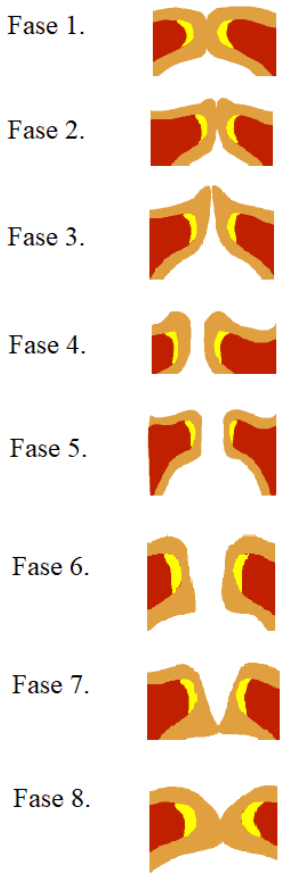


Figura 1.5. Ciclo fonatorio.
Adaptado de [10].

Tras el paso por la cavidad laríngea el aire llega al tracto vocal, que actúa como una caja de resonancia, aquí, dependiendo de la posición de los órganos articulatorios, se generan distintos sonidos. Las resonancias generadas tienen su energía focalizada alrededor de determinadas frecuencias del espectro, conocidas como formantes [6].

1.2. LOS SONIDOS

1.2.1. Clasificación del sonido

En el habla se pueden distinguir, de acuerdo con su articulación y distribución, dos tipos de sonidos, vocales y consonantes. Para la producción de las consonantes, los órganos articuladores crean una obstrucción al flujo glotal, mientras que para la producción de las vocales dicha obstrucción no existe [11].

Vocales

Los sonidos vocales se pueden clasificar de acuerdo con la posición de la lengua en [2]:

- Abierta: la lengua está completamente separada del paladar [a][1].
- Media: la lengua se encuentra a una distancia media del paladar [e,o].
- Cerrada: la lengua está muy cerca del paladar [i,u].

Consonantes

Para la generación de las consonantes un órgano articulatorio activo, como por ejemplo el labio inferior, la punta de la lengua o el dorso de la lengua, se mueve para hacer contacto con un órgano articulatorio pasivo, como por ejemplo los dientes o el paladar.

En la clasificación de las consonantes se tienen en cuenta tres parámetros: el punto de articulación, el modo de articulación y la función de las cuerdas vocales [11].

Punto de articulación

La clasificación según el punto de articulación depende del lugar en el que se restringe el flujo del aire [11], como se muestra en la Figura 1.4.

- Bilabial: hay contacto entre el labio superior e inferior [p, b, β, m].
- Labiodental: el órgano articulador activo es el labio inferior, el cual se desplaza para hacer contacto con el articulador pasivo, que en este caso es el borde de los dientes superiores [f, v, ɱ].
- Dental: hay contacto entre el ápice de la lengua y los incisivos superiores [t, d, ð].
- Interdental: la lengua se ubica entre los dientes superiores e inferiores [θ].
- Alveolar: la punta de la lengua está en contacto con la región alveolar [n, l, ɾ, r̄, s, z].
- Prepalatal: hay contacto entre la parte delantera del dorso de la lengua y una parte extensa de la parte anterior de la boca, entre la zona alveolar y el paladar [ʃ, ʧ, ʒ, ʤ].
- Palatal: el ápice de la lengua se eleva hacia el paladar [ʝ, ɟ, ɲ, ʎ].
- Velar: la parte detrás del dorso de la lengua está en contacto con el velo del paladar [k, g, ɣ, x, ŋ].
- Glotal: la obstrucción tiene lugar en la glotis [h, ɦ].

[1] Los símbolos entre corchetes corresponden a la notación fonética de los sonidos generados en cada uno de los casos.

Modo de articulación

Se refiere al tipo de obstrucción que se crea en la articulación de las consonantes [11]**.**

- Oclusiva: los órganos articuladores hacen un contacto firme, de modo que se detiene totalmente el flujo del aire. Se produce en dos etapas, el cierre del tracto seguido de una apertura súbita [p, t, k].
- Fricativa: no se interrumpe completamente el paso del aire, éste se escapa a través de una constricción entre los articuladores, lo que produce una turbulencia [f, s, x].
- Aproximante: el articulador activo solo se aproxima al articulador pasivo, lo que no genera fricción o turbulencia [β, ð, ɣ].
- Africada: La articulación incluye dos etapas: oclusión y suelte fricativo. En español solo hay un fonema africado [ʧ].
- Nasal: Se produce con la oclusión completa del canal oral, sin embargo, permite el flujo de aire a través del canal nasal [m, n, ŋ].
- Vibrante: En español existen dos fonemas vibrantes, la vibrante simple [ɾ], que se produce con un contacto rápido del ápice con la zona alveolar y la vibrante múltiple [r̄], que se origina cuando hay dos o más contactos rápidos.
- Lateral: Hay contacto entre los articuladores en el centro de la cavidad oral y el aire circula a través de uno o ambos lados [ʎ].

Función de las cuerdas vocales

- Sonoras: Las cuerdas vocales se unen incrementando la presión subglotal, lo que origina que se separen al salir el aire de los pulmones. Cuando disminuye la presión del aire, sin embargo, otras fuerzas intentan unir las cuerdas vocales de nuevo, lo que ocasiona un breve proceso de vibración, como resultado se produce un sonido sonoro [ʃ, f, x].
- Sordas: Si las cuerdas vocales están totalmente separadas, no se produce una vibración con el paso del aire, por lo que el resultado es un sonido sordo [ʒ, v, ɣ].

1.3. PRINCIPALES CARACTERÍTICAS DEL ANÁLISIS ACÚSTICO

Existen 2 características fundamentales que se relacionan con la señal acústica, en el tiempo y en la frecuencia, éstas son la frecuencia fundamental y la intensidad:

- *Frecuencia fundamental o pitch.* Es la frecuencia más baja del espectro de frecuencias de la señal de voz y se representa con f_0 . Equivale al número de veces que las cuerdas vocales se abren y se cierran por segundo, y se

expresa en ciclos por segundo o Hercios (Hz). La laringe humana puede producir un conjunto amplio de frecuencias, denominado rango vocal, que varía en función de la edad y el sexo [12]. La frecuencia fundamental de los hombres normalmente varía entre los 80 y 150 Hz y la de las mujeres entre los 140 a 250 Hz. Por esta razón se puede decir que las mujeres tienen voces más agudas que las de los hombres, pues su frecuencia fundamental es más elevada [5]. El efecto psicoacústico de la frecuencia es el tono vocal, no obstante, es importante resaltar que el tono no solo depende de f_0, sino también de parámetros como la intensidad o la composición espectral, sin embargo, estos últimos juegan un papel secundario. Al aumentar f_0, el tono se hace más agudo, del mismo modo al disminuir f_0 el tono se hace más grave. Estas variaciones no son lineales y no se percibe igual el aumento de una frecuencia baja que el de una frecuencia alta [12].

- *Envolvente Espectral.* Es una curva en el dominio de la frecuencia que se deriva del espectro de magnitud de una señal de voz, i.e., contiene los cambios lentos del espectro de magnitud. Gracias a esto se pueden identificar los formantes de dicho espectro, los cuales proporcionan información acerca del sonido que se está pronunciando.

1.4. CONSIDERACIONES PARA EL PROCESAMIENTO DE SEÑALES DE VOZ

La vibración de las cuerdas vocales genera una señal periódica, la cual en el dominio de la frecuencia tiene un comportamiento discreto caracterizado por los armónicos, dichos armónicos son componentes de frecuencia separadas entre ellas por múltiplos enteros de f_0. La magnitud de estas componentes tiene una pérdida de 12 dB por cada octava que aumenta la frecuencia (ver Figura 1.6).

A partir del espectro mostrado en la Figura 1.6 no es posible identificar un sonido vocal primario[2], e.g. una vocal definida, debido a que para generar el sonido la señal debe viajar a través de las cavidades supraglóticas, donde sufre modificaciones debidas a la resonancia.

La resonancia cambia la relación de amplitudes de los armónicos ya que estos pueden verse atenuados o amplificados y, adicionalmente, crea nuevas componentes en frecuencia; sin embargo, es posible establecer unos máximos relativos de amplitud en el espectro, los cuales se denominan *formantes* [12]. En esta etapa de la producción de la voz el sonido vocal primario toma una estructura

[2] Es un sonido de una lengua natural hablada que se pronuncia con el tracto vocal abierto, no habiendo un aumento de la presión del aire en ningún punto más arriba de la glotis.

formántica (ver Figura 1.7), en este punto ya es posible realizar la distinción de las distintas vocales.

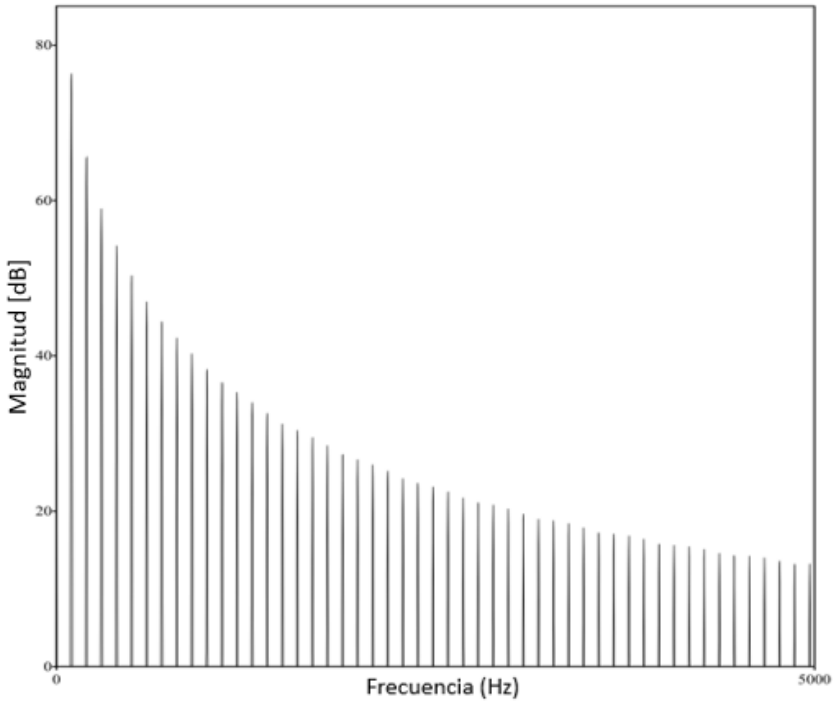


Figura 1.6. Espectro de la fuente glotal.
Adaptado de [12].

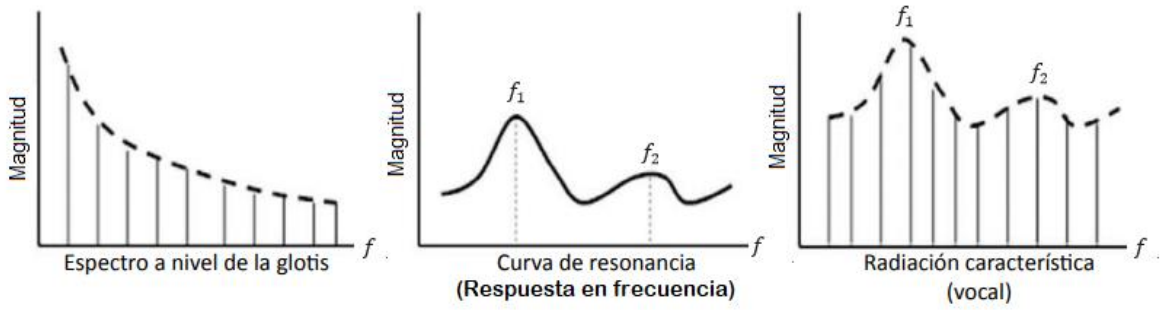


Figura 1.7. Efecto de la curva de resonancia particular del tracto vocal.
Adaptado de [12].

El proceso de producción de la voz se puede sintetizar mediante un modelo fuente-filtro. La fuente es la señal que corresponde a la perturbación acústica periódica generada por la corriente de aire procedente de los pulmones, esta señal es modificada al atravesar las cavidades supraglóticas, i.e. tractos vocal y nasal, las cuales se comportan como un filtro. La respuesta en frecuencia del filtro equivale al conjunto de formantes que caracterizan la curva de resonancia, como se muestra en la Figura 1.7. Un formante es un modo de vibración natural del tracto vocal, está compuesto por una frecuencia central o también llamada frecuencia formantes y un

ancho de banda. Las frecuencias formantes más significativas del tracto supraglótico, en cuanto a la acústica de la voz hablada, son las primeras cuatro (f_1, f_2, f_3 y f_4), donde las dos primeras, f_1 y f_2, son las menos agudas y permiten la identificación de las vocales. La señal de salida está conformada por las mismas componentes de frecuencia, pero con amplitudes moduladas por el filtro [12] (ver Figura 1.8).

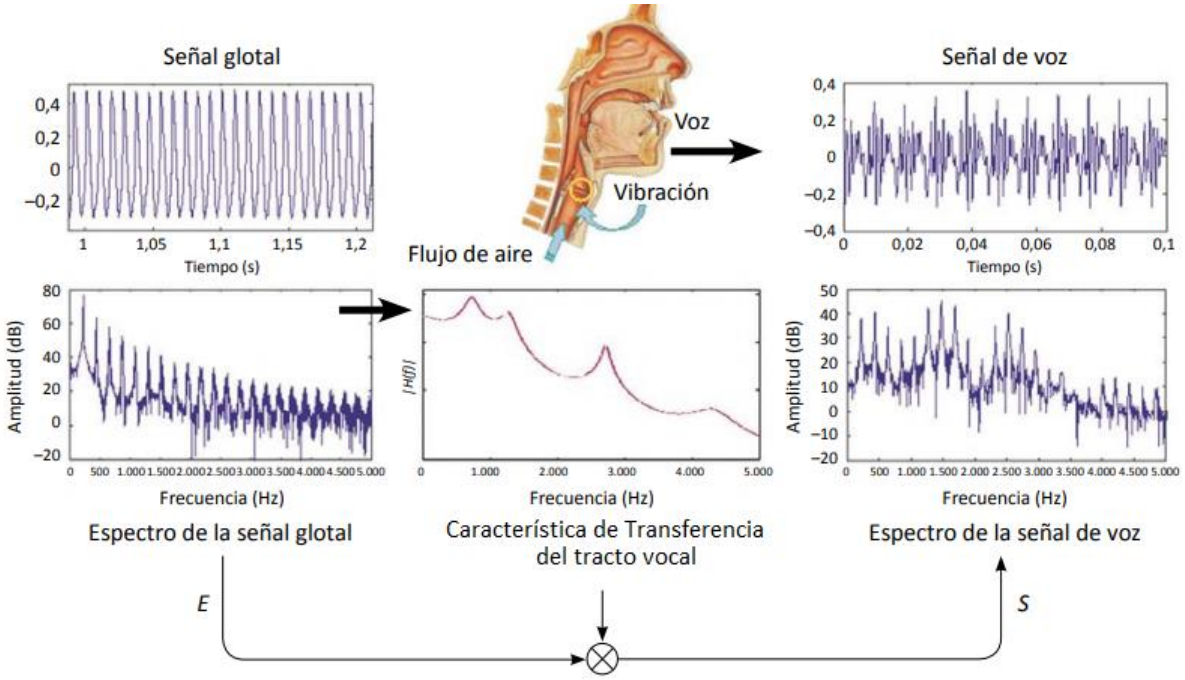


Figura 1.8. Esquema de producción de la voz.
Adaptado de [12].

Como se mencionó en la sección 1.2 los sonidos se pueden clasificar, según la función de las cuerdas vocales en sonoros y sordos. Los sonidos sonoros son aquellos que se producen por medio de la vibración de las cuerdas vocales, entre ellos se encuentran las vocales y algunas consonantes como d, b, l, m, entre otras. En la Figura 1.9 se muestra el espectro de un sonido o voz sonora, donde, los máximos de la envolvente muestran tres formantes (f_1, f_2, f_3) y f_0. Además, es destacable que, si la señal fuese periódica, el espectro representado sería discontinuo, pero como no lo es totalmente, se dice que es una señal cuasi periódica.

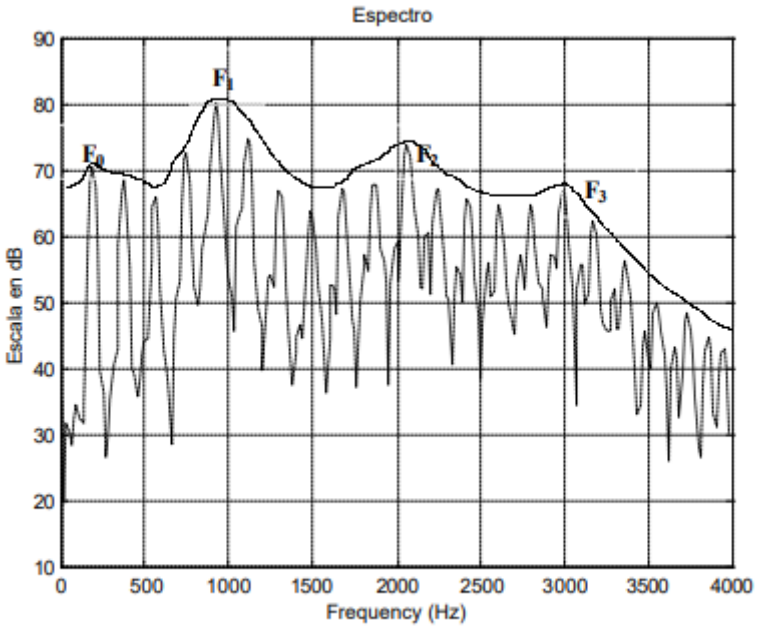


Figura 1.9. Espectro de una señal sonora.
Adaptado de [13].

Los sonidos sordos, o la voz sorda, no se generan por la vibración de las cuerdas vocales. En este caso, las cuerdas vocales se encuentran separadas generando una constricción del tracto vocal, entonces, el aire pasa por dicha constricción a una velocidad tal que se genera una corriente de turbulencia, y como resultado, la señal de voz resultante es de comportamiento aleatorio en forma de ruido blanco o gaussiano. Como ejemplos de voz sorda se encuentran las consonantes ‘s’, ‘z’, ‘f’; entre otras.

Es importante tener en cuenta que las características de las señales de voz cambian constantemente para reflejar los sonidos que se están pronunciando (sonoros y no sonoros), como se observa en la Figura 1.10, por tanto, la voz no es estacionaria; no obstante, si se examina en intervalos lo suficientemente cortos (entre 5 y 100 ms) es posible asumir un comportamiento aproximadamente estacionario.

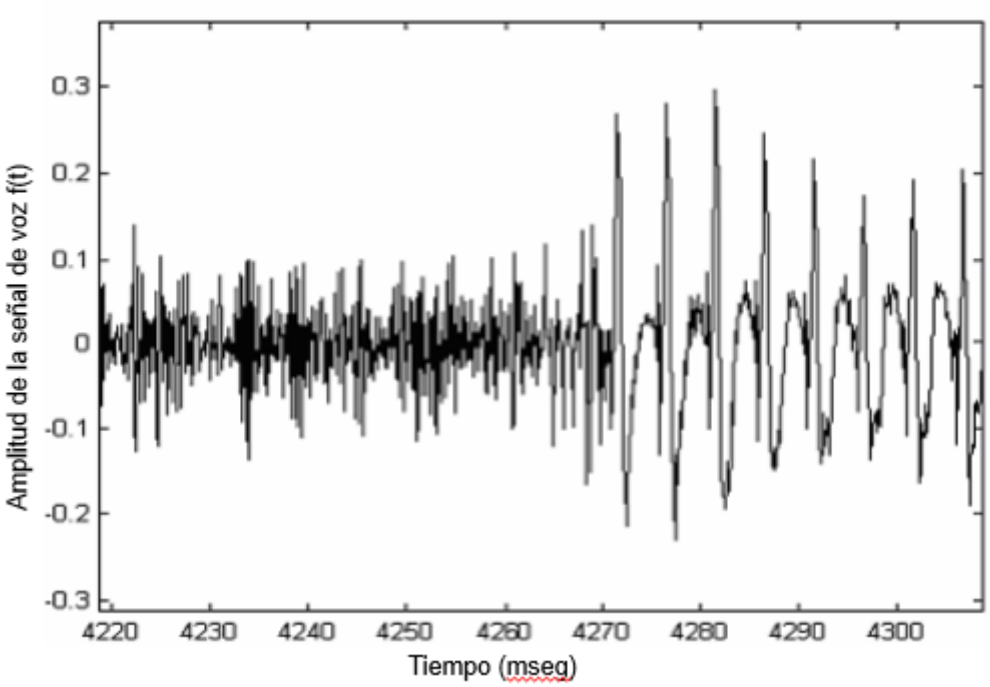


Figura 1.10. Transición de un sonido sordo a un sonido sonoro.
Adaptado de [13].

1.5. ANCHO DE BANDA DE LA SEÑAL DE VOZ

El espectro de una señal de voz decae rápidamente (ver Figura 1.11), por lo que su ancho de banda (medido en Hercios, Hz) se calcula como la diferencia entre las frecuencias mínimas y máximas presentes en la señal que superen cierto umbral, dicho umbral busca anteponer las componentes más significativas para la representación de la señal.

El valor del umbral conduce a un determinado ancho de banda, dado que las señales de voz no son finitas en el dominio de la frecuencia y por tanto se deben limitar. En general se asume que el espectro de la voz humana se extiende aproximadamente desde los 50 Hz a los 8 KHz [14]. En la Figura 1.11 se muestran los anchos de banda considerados en el procesamiento de señales de voz, en donde se observa que un mayor ancho de banda conlleva a una mejor representación de los sonidos, y por lo tanto a una mayor calidad.

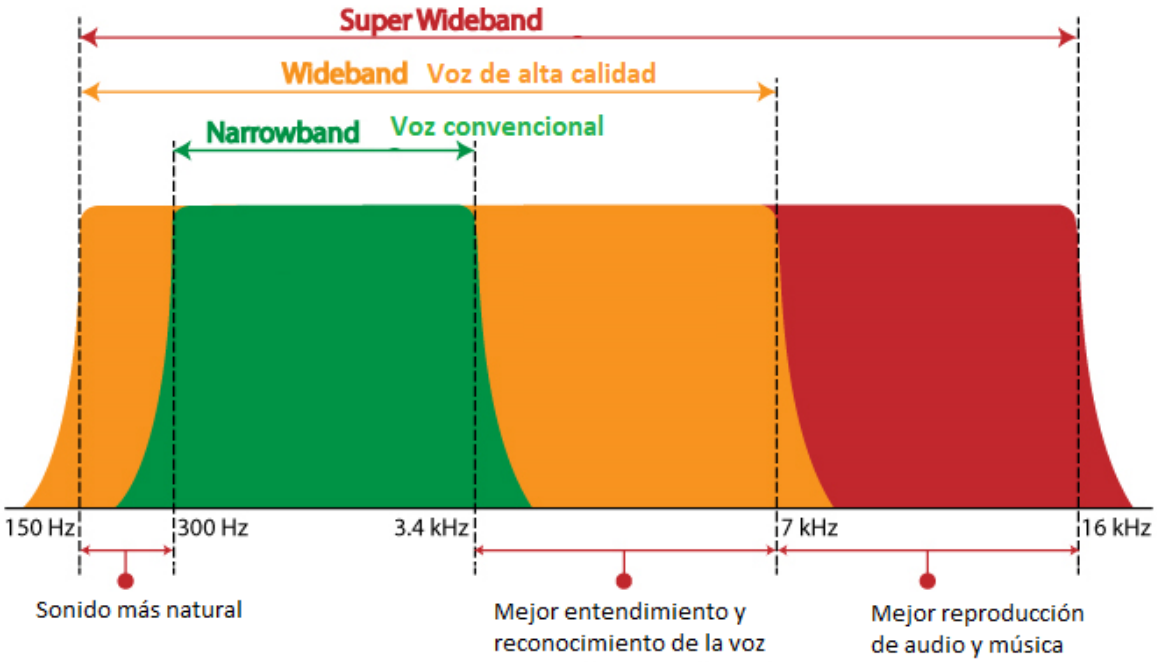


Figura 1.11. Bandas de frecuencias de la voz.
Adaptado de [15].

CAPÍTULO 2: *VOICE MORPHING*

En el presente capítulo se aborda el concepto de *Voice Morphing*, el cual es clave para el desarrollo de este trabajo de grado. En la sección 2.1 se describe la definición de *Voice Morphing,* luego, en la sección 2.2 se presentan las aplicaciones de *Voice Morphing* en diversas áreas. En la sección 2.3 se hace una descripción general de un sistema de *Voice Morphing* y algunas técnicas utilizadas para su implementación. Finalmente, se encamina la investigación hacia los sistemas basados en transformaciones lineales en la sección 2.4.

2.1. DEFINICIÓN DE *VOICE MORPHING*

Voice Morphing es una técnica relativamente nueva que abarca un amplio campo investigativo, además de estar muy relacionada con otras técnicas, como la conversión y la transformación de la voz, por lo cual, existen diversas variaciones en su definición, en el presente trabajo de grado se opta por trabajar con una definición general, dentro de la cual se incluyen como casos particulares las otras definiciones. *Voice Morphing* es un procedimiento para hacer que una voz determinada suene como una voz diferente [16], mediante esta técnica se modifica algorítmicamente la voz de un hablante original para que suene como si hubiese sido pronunciada por un hablante designado, conservando siempre el mensaje. La voz original y la voz designada se denominan voz origen y objetivo, respectivamente.

Existen tres tipos de *Voice Morphing* a saber:

- *Voice Morphing* con una referencia: en este proceso se utiliza una voz origen para generar una nueva voz modificada. Se toman las características de la voz origen y se modifican para sintetizar una nueva voz denominada voz objetivo [16]. Este tipo de *Voice Morphing* se relaciona ampliamente con la conversión de voz y es el enfoque principal de este trabajo.

- *Voice Morphing* con dos referencias: consiste en el uso de dos voces de referencia u origen para generar una voz objetivo. En este tipo de *Voice Morphing* se extraen determinadas características de las voces origen y se mezclan para generar una tercera voz [16].

- *Voice Morphing* generalizado: consiste en la mezcla de muchas voces origen para generar una nueva voz, permitiendo realizar combinaciones específicas para obtener la voz objetivo. En este orden de ideas, *Voice Morphing* con dos referencias es un caso especial de Voice Morphing generalizado [16].

2.2. APLICACIONES DE *VOICE MORPHING*

Inicialmente *Voice Morphing* era utilizada únicamente en estudios de doblaje; sin embargo, actualmente tiene otras aplicaciones como la Traducción de Habla a Habla (SST, *Speech to Speech Translation*) [17], la personalización de las voces para los sistemas de Síntesis de Texto a Habla (TTS, *Text to Speech Synthesis*) a un menor costo, y aplicaciones de voz con lectores de pantalla y correo electrónico para personas con discapacidad visual.

Existe, además, una aplicación significativa en el campo del entretenimiento, donde voces sintetizadas y con características específicas son requeridas, como en el caso de los videojuegos y voces en *off*[3] para anuncios y películas [18].

El área de la salud también es un gran campo de aplicación, ya que en ésta se busca hacer comprensible la voz de personas con problemas en el sistema fonador, como es el caso de los laringectomizados [19], los cuales son pacientes a quienes les han practicado procedimientos quirúrgicos en la laringe.

El disfraz de voz y la codificación de voz son otras de las aplicaciones, donde la transmisión de la voz puede hacerse sin exponer la identidad del hablante [20]. Es importante mencionar que el disfraz de la voz representa una herramienta para los predadores, quienes pueden utilizarlo para engañar a personas haciéndose pasar como un individuo de menor edad o del sexo opuesto e incluso suplantar identidades.

2.3. SISTEMA DE VOICE MORPHING

Para construir un sistema de *Voice Morphing* existen tres aspectos interdependientes que son necesarios analizar.

I. Elección de un modelo matemático para la descomposición y regeneración de la señal de voz [21]. El modelo sinusoidal, de acuerdo con varias investigaciones [19], [22], [23], [24], es uno de los modelos más apropiados, puesto que permite la manipulación de la señal de voz, sin introducir artefactos[4] significativos.

[3] Voz pregrabada y luego añadida en el proceso de producción que pertenece a un individuo que no está visible frente a la cámara, se usa frecuentemente en documentales con una función descriptiva.
[4] Alteraciones en la señal de la voz que afectan la naturalidad.

II. Extracción y codificación de las características acústicas, independientes del mensaje y del entorno, que identifican al hablante origen y al hablante objetivo [19], así como también, determinar el tipo de función de transformación y el método de entrenamiento del modelo de conversión [21].

III. Conversión de las características del hablante origen, a través de la función de transformación aplicando el modelo de conversión establecido.

Los sistemas de *Voice Morphing* generalmente están compuestos por dos etapas (ver Figura 2.1), etapa de entrenamiento y etapa de transformación. La etapa de entrenamiento se enfoca en los dos primeros puntos, anteriormente mencionados, mientras que la etapa de transformación se orienta principalmente hacia el tercer punto.

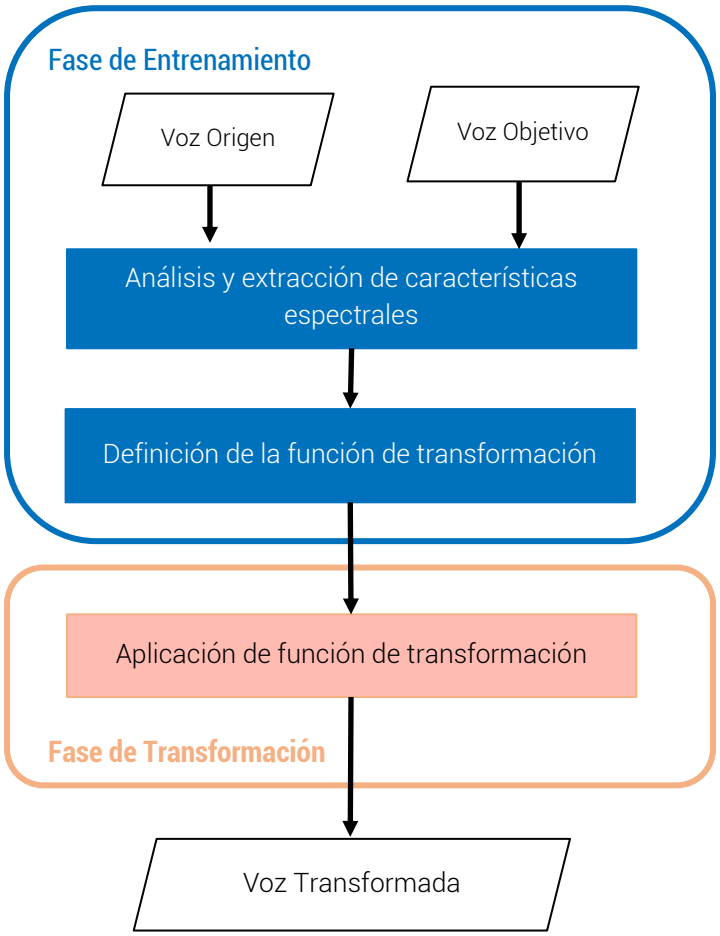


Figura 2.1. Diagrama de bloques de un sistema de *Voice Morphing*.
Por los autores.

2.3.1. Etapa de entrenamiento

En la etapa de entrenamiento se analizan las señales de voz del hablante origen y del hablante objetivo, para lo cual se utilizan muestras de voz de entrenamiento, con el objetivo de crear una transformación que mapee el espacio que caracteriza el habla del hablante origen con el del hablante objetivo.

Voice morphing tiene como objetivo el control de la identidad del hablante. Identidad que está delimitada por el tono medio[5], la estructura del formante y las características del tracto vocal; éstas dos últimas pueden representarse mediante la forma general de la envolvente espectral, característica ampliamente utilizada en la mayoría de los sistemas de *Voice Morphing* [21].

La extracción de las características, normalmente las frecuencias formantes, de dichas señales se efectúa usando Codificación de Predicción Lineal (LPC, *Linear Prediction Coding*) [20], Coeficientes Cepstrales en las Frecuencias de Mel (MFCC, *Mel Frequency Cepstral Coefficients*) o Frecuencias Espectrales de Línea (LSF, *Line Spectral Frequencies*) [19]. Otros métodos utilizados se basan en procedimientos mixtos en el dominio del tiempo y de la frecuencia [18].

En esta etapa también se define la función de transformación y el método de entrenamiento del modelo para la conversión.

Se han planteado, en diversas investigaciones, varios enfoques para el diseño de la función de transformación; entre los que se encuentran: el mapeo de libros de códigos, las Redes Neuronales Artificiales (ANN, *Artificial Neural Networks*) [25], las Funciones de Base Radial (RBF, *Radial Basis Functions*) [26] y enfoques basados en la modificación del *pitch* a partir de múltiples segmentos del habla traslapados (PSOLA, *Pitch Synchronous Overlap and Add*) [27].

2.3.2. Etapa de Transformación

En esta etapa se aplica la función de transformación, determinada en la etapa de entrenamiento, a las características del hablante origen.

Aunque se han desarrollado sistemas que reducen distorsiones introducidas durante la transformación, se siguen presentando falencias considerables. Actualmente existe un amplio campo de investigación enfocado en implementar soluciones, para

[5] Promedio de las diferentes frecuencias fundamentales que se pueden detectar en una muestra de voz debido a que la vibración de las cuerdas vocales no es constante a lo largo de un discurso.

garantizar la calidad de la voz sintetizada, entre las cuales se destacan los sistemas basados en transformaciones lineales [19].

2.4. SISTEMAS BASADOS EN TRANSFORMACIONES LINEALES

El enfoque de este trabajo de grado está dirigido hacia los sistemas de *Voice Morphing* que hacen uso de las transformaciones lineales. Por lo tanto, es de gran importancia introducir el esquema general de estos sistemas. En la Figura 2.2 se presenta un diagrama, que se alinea al sistema de *Voice Morphing* general, propuesto anteriormente, el cual expone el sistema de *Voice Morphing* estructurado en las dos etapas generales (entrenamiento y transformación), dentro de las cuales se especifica en mayor grado los procedimientos y técnicas utilizadas.

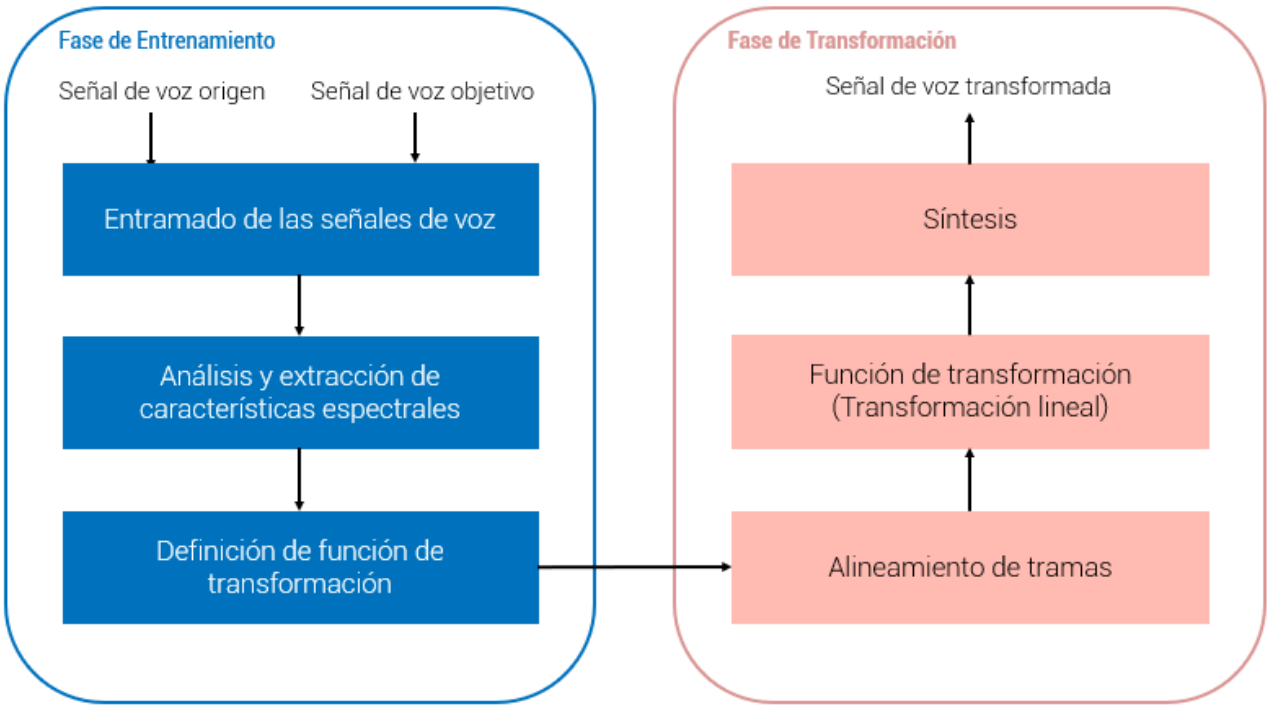


Figura 2.2. Diagrama de bloques de un sistema de *Voice Morphing* basado en transformaciones lineales. Por los Autores.

2.4.1. Etapa de entrenamiento de un sistema basado en Transformaciones Lineales

Para extraer las características y representar la envolvente espectral de las señales de voz del hablante origen y del hablante objetivo, en los sistemas que usan transformaciones lineales, se utilizan generalmente los MFCC, debido a su bajo costo computacional y a su robustez [28]. Además, para la definición de la función

de transformación, se aplica un Modelo de Mezcla Gaussiana (GMM, *Gaussian Mixture Model*), que es un método de aprendizaje no supervisado que permite caracterizar la estructura de los formantes en la envolvente espectral de la señal de voz sin introducir artefactos significativos.

Codificación de Predicción Lineal y Frecuencias Espectrales de Línea

La codificación de predicción lineal es una de las principales técnicas de análisis del habla. En LPC, se estima los polos de los filtros para los segmentos de voz de corta duración. La muestra actual s_n se predice como una combinación lineal de sus p muestras pasadas

$$s_n = \sum_{k=1}^{p} a_k s_{n-k}, \tag{2.1}$$

donde a_k es el k-ésimo coeficiente del filtro de todos los polos $A(z)$. Los coeficientes del filtro de predicción lineal se pueden calcular de a través de distintos métodos. El método de autocorrelación con recursividad de Levinson-Durbin es uno de los más utilizados en el cálculo de estos coeficientes [29].

Los coeficientes LPC se pueden transformar en frecuencias espectrales de línea, conservando una conversión completamente reversible. Las LSF se obtienen calculando las raíces de dos polinomios, $P(z)$ (Ecuación 2.2) y $Q(z)$ (Ecuación 2.3).

$$P(z) = A(z) + z^{-(P+1)}A(z^{-1}) \tag{2.2}$$

$$Q(z) = A(z) - z^{-(P+1)}A(z^{-1}) \tag{2.3}$$

Los parámetros LSF ofrecen una representación robusta para objetivos de cuantificación y modificación. Debido a que estos parámetros tienen una estrecha relación con los formantes se han convertido en técnicas de caracterización populares en los sistemas de análisis de habla [29]. En el Apéndice A se hace un análisis más detallado de las LSF.

Coeficientes Cepstrales en las Frecuencias de Mel

Los MFCC son una de las técnicas de extracción de características más utilizada en reconocimiento del habla, debido a que representan la envolvente espectral de la señal de voz en forma compacta [30].

El concepto de coeficientes MFCC utiliza una escala de frecuencia no lineal denominada Mel para imitar el comportamiento psicoacústico a tonos puros de distinta frecuencia del oído humano. Distintos estudios han demostrado que el sistema auditivo humano procesa la señal de voz en el dominio espectral y se caracteriza por tener mayores resoluciones en bajas frecuencias, lo que se logra por medio de la escala Mel. De esta forma se asigna mayor relevancia a las bajas frecuencias de forma análoga a como se hace en el sistema auditivo humano [31].

En la Figura 2.3 se muestra el diagrama de bloques del proceso para la extracción de los vectores característicos MFCC.

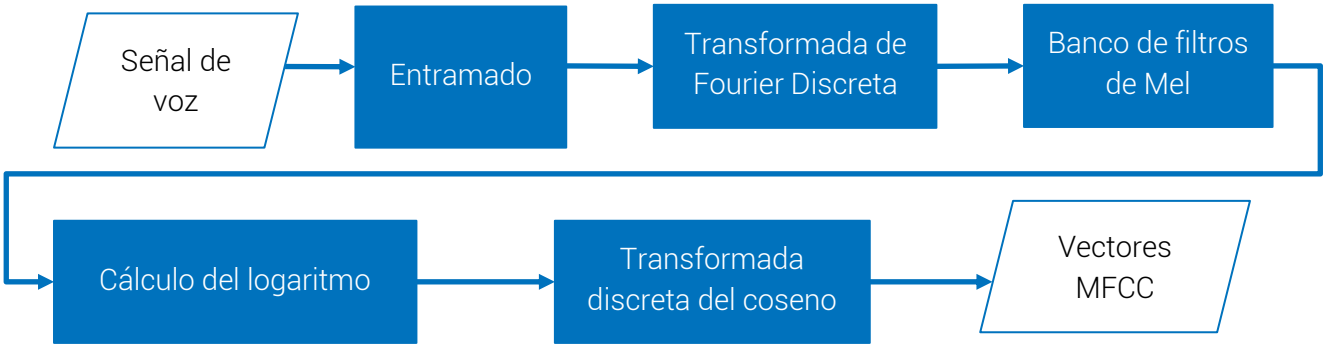


Figura 2.3. Diagrama de bloques para la extracción de coeficientes MFCC.
Por los autores.

El primer paso es dividir la señal de voz en tramas, usualmente aplicando una función de ventaneo. El ventaneo sirve para eliminar los bordes de la señal y darle una acentuación a la parte central de la trama para su análisis [28]. En este proceso se generan tramas o segmentos consecutivos de la señal [31].

A continuación, se calcula la Transformada Discreta de Fourier (DFT, *Discrete Fourier Transform*) de tamaño N de las tramas aplicando la Ecuación 2.4.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk}, 0 \leq k \leq N. \quad (2.4)$$

En esta etapa se descarta la información de fase, y la magnitud de la señal $X[k]$ se multiplica por un banco de F filtros triangulares, los cuales se encuentran espaciados de acuerdo con la escala de frecuencias Mel (Ver Figura 2.4).

$$mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right), \quad (2.5)$$

donde f es la frecuencia en escala lineal [Hz] [29].

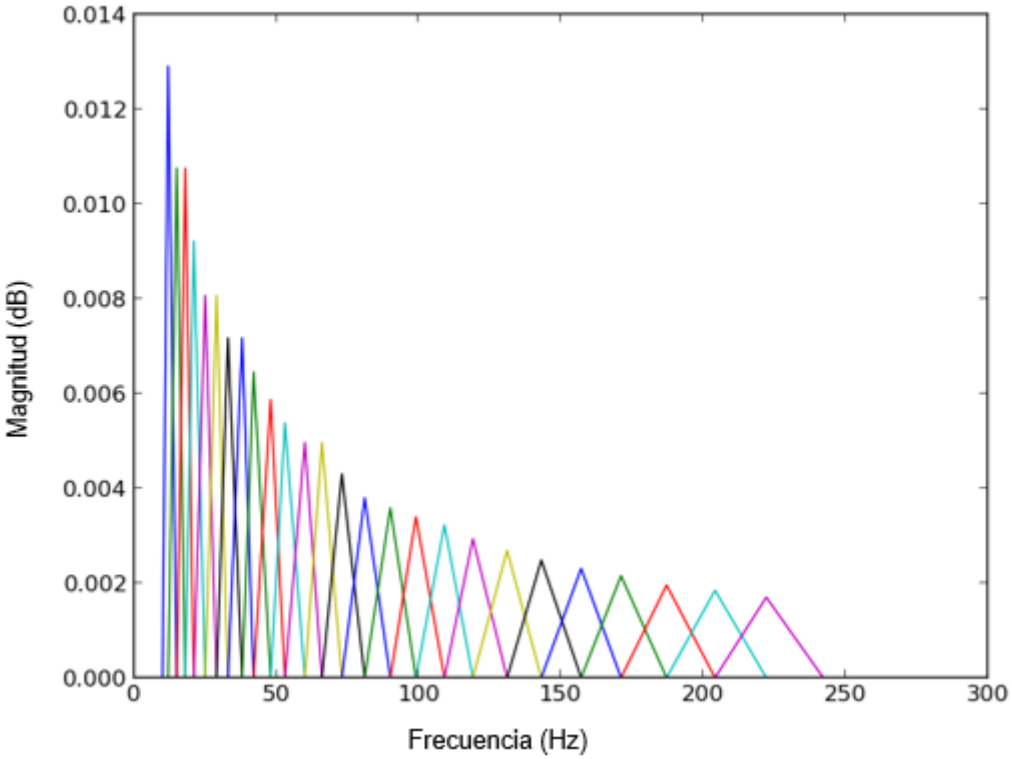


Figura 2.4. Banco de filtros Mel.
Adaptado de [32].

Después de que la envolvente de la señal es multiplicada por el banco de filtros, se estima la energía en cada uno de los filtros (Ecuación 2.6).

$$E_m = \sum_{k=0}^{N-1} |X[k]|^2 H_m[k]\,, 1 \le m \le F. \tag{2.6}$$

A continuación, se calcula el logaritmo, pasando al dominio de la potencia espectral logarítmica y finalmente se aplica la Transformada Discreta del Coseno (DCT, *Discrete Cosine Transform*), la cual lleva los coeficientes espectrales al dominio de la frecuencia, convirtiéndolos en los coeficientes MFCC [31]. De este vector obtenido se toman la cantidad de coeficientes deseados por trama [28].

$$C_{MFCC}[m] = \sum_{k=0}^{N-1} \log(E_k) \cos\left(m\left(k - \frac{1}{2}\right)\frac{\pi}{N}\right), \quad m = 1, \dots, F. \tag{2.7}$$

2.4.2. Etapa de Transformación de un Sistema Basado en Transformaciones Lineales

A partir de los coeficientes calculados como se muestra en la sección 2.4.1 es posible estimar la envolvente espectral de la señal de voz, como se muestra en la Figura 2.5, no obstante, ante un mismo número de coeficientes, se obtienen envolventes con evidentes diferencias. Dado que para la transformación de las señales de voz se utilizan los modelos de mezcla gaussiana, se selecciona la envolvente espectral que resulta de los MFCC, debido a que su comportamiento es más suave.

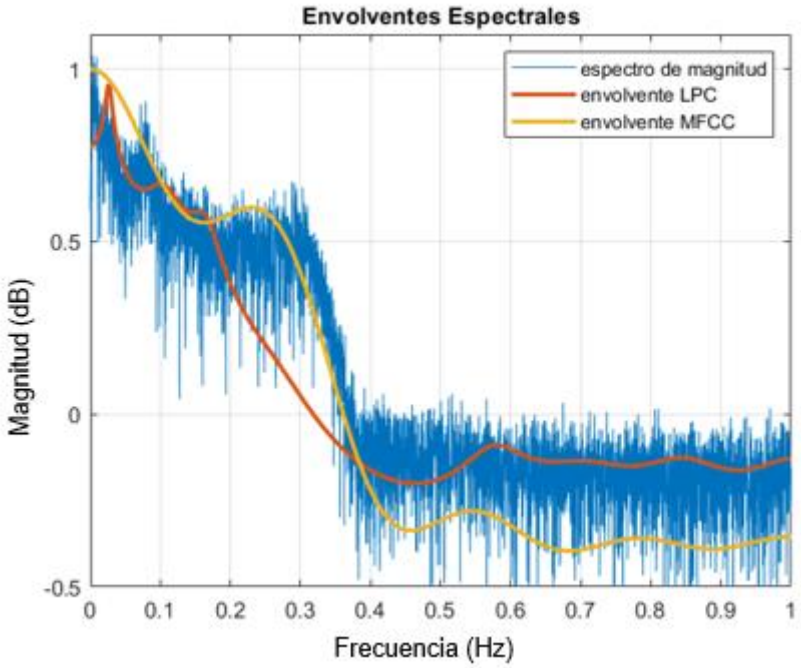


Figura 2.5. Envolventes espectrales.
Por los autores.

Modelos de Mezcla Gaussiana para la caracterización de señales de voz

Un Modelo de Mezcla Gaussiana (GMM, *Gaussian Mixture Model*) es una Función de Densidad de Probabilidad (PDF, *Probability Density Function*) paramétrica representada como una suma ponderada de PDF normales.

Tanto dentro como fuera del dominio del reconocimiento de voz, los GMM se usan generalmente para la caracterización y clasificación de un conjunto de datos. Los GMM son bien conocidos por su capacidad para representar distribuciones arbitrariamente complejas con múltiples modos [33] y como modelo para estimar las

características de la envolvente espectral de una señal de voz basada en tramas [34].

La técnica asume que un conjunto de componentes gaussianos puede representar una distribución, $p(\mathrm{x})$, la cual se basa en las características de la envolvente espectral de la señal de voz. Un GMM es una suma ponderada de M componentes gaussianos, como se muestra en la Ecuación 2.8.

$$p(\mathrm{x}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\mathrm{x}; \mu_m^{(x)}, \Sigma_m^{(x)}\right), \tag{2.8}$$

donde α_m es la probabilidad a priori del m-ésimo componente gaussiano, $0 \leq \alpha_m \leq 1$, y el término $\mathcal{N}(\mathrm{x}; \mu_m^{(x)}, \Sigma_m^{(x)})$ denota la distribución normal multivariante con el vector de medias μ_m y la matriz de covarianza Σ_m [29]. Es importante resaltar que se debe cumplir que

$$\sum_{m=1}^{M} \alpha_m = 1.$$

En este caso x son los MFCC que parametrizan la envolvente espectral normalizada (ver Apéndice B), y corresponden a las observaciones que sustentan la estimación de los parámetros de GMM. La estimación se realiza iterativamente, utilizando el algoritmo Expectativa-Maximización (EM, *Expectation-Maximization*).

EM trabaja alternando entre 2 pasos denominados E (Expectativa) y M (Maximización). El paso E busca calcular los pesos de cada uno de los componentes gaussianos, a partir de los cuales el paso M adapta los parámetros del modelo [34]. En la Figura 2.6 se muestra cómo se adapta el algoritmo al aumentar el número de iteraciones, inicialmente las medias de las gaussianas toman valores aleatorios y sus parámetros se ajustan a partir de los valores de x, por lo cual se basa en la Máxima Verosimilitud (ML, *Maximum Likelihood*).

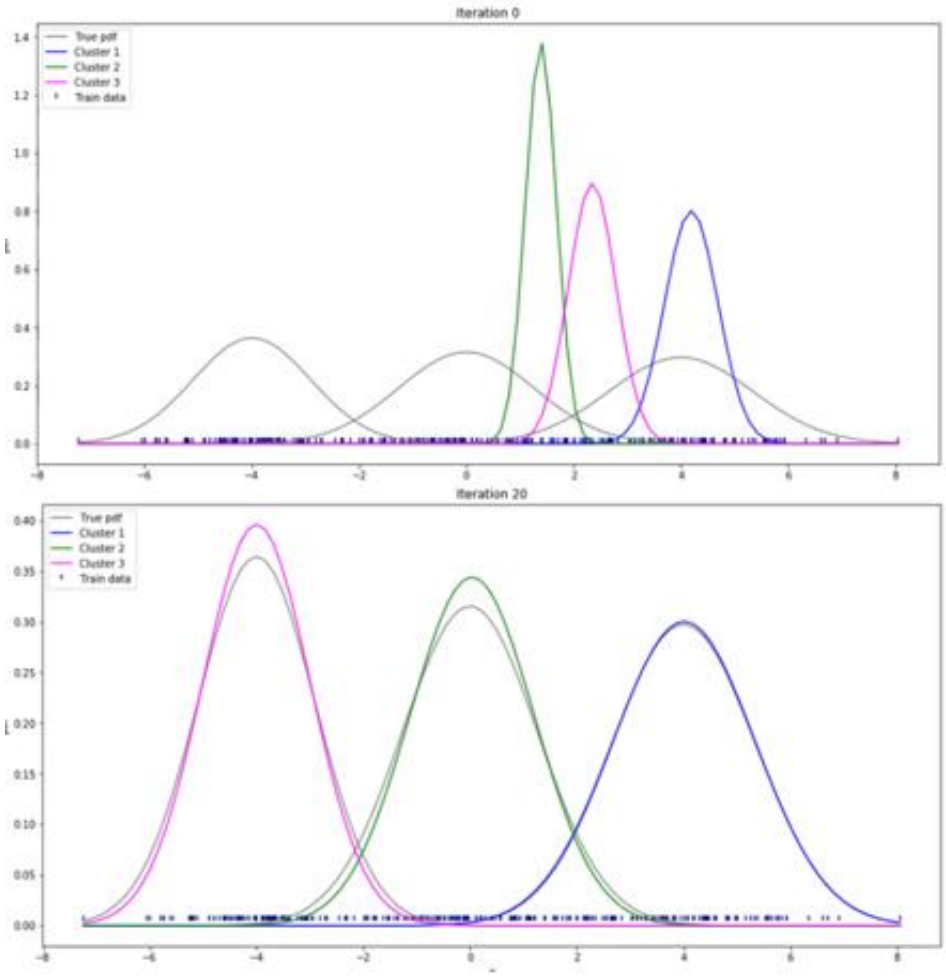


Figura 2.6. Estimación espectral de GMM.
Adaptado de [35].

Las medias, las varianzas y los pesos de mezcla de las funciones de densidad de probabilidad representan las frecuencias de los formantes, los anchos de banda y las amplitudes, respectivamente. La Figura 2.7 muestra una distribución de mezcla estimada de tres componentes gaussianos superpuestos sobre el espectro de magnitud de DFT, que se obtiene mediante el análisis de un segmento corto del habla [36].

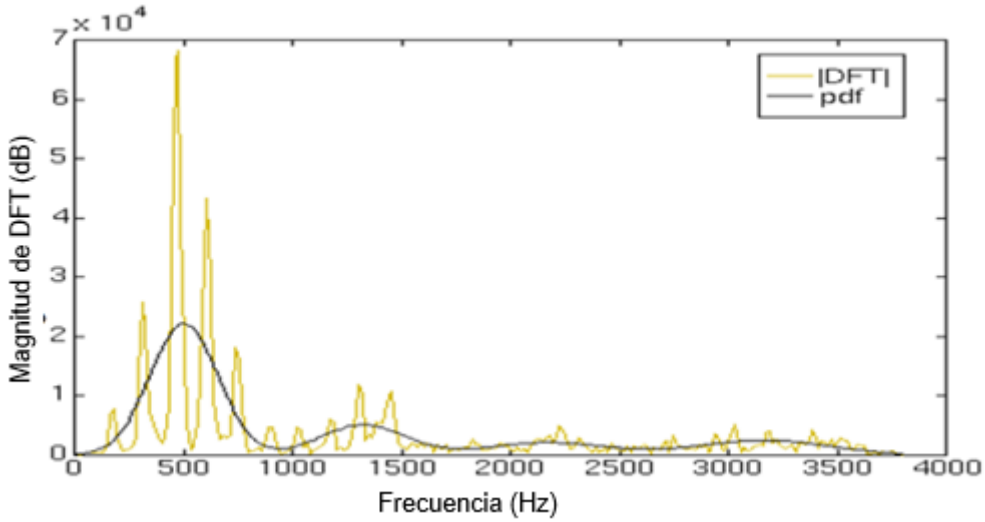


Figura 2.7. Mezcla de gaussianos ajustados a un espectro de magnitud DFT.
Adaptado de [36].

Como puede verse, la forma espectral ha sido bien representada y los formantes dentro del rango de frecuencia han sido seleccionados por cada uno de los componentes gaussianos en la mezcla.

Posterior a la extracción y representación de las características de la envolvente espectral de las señales del hablante origen y del hablante objetivo, se entrena una función de conversión para capturar la correlación entre las características de las señales de voz origen y objetivo [19]. Esta función está basada en transformaciones lineales.

Transformaciones lineales

Se tiene dos conjuntos de vectores espectrales $\mathbf{X}$ y $\mathbf{\Psi}$ que codifican respectivamente las características de la envolvente espectral de las señales del hablante origen y del hablante objetivo:

$$\mathbf{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_i];\ \mathbf{\Psi} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_i]; \qquad (2.9)$$

donde cada vector $\boldsymbol{X}_n$ (o $\boldsymbol{Y}_n$) es de dimensión p *y los conjuntos de vectores* $\mathbf{X}$ *y* $\mathbf{\Psi}$ *de* longitud i. Lo que se desea es una función $F()$ tal que la envolvente transformada $F(\boldsymbol{X}_n)$ coincida mejor con la envolvente de destino $\boldsymbol{Y}_n$ para todas las envolventes del conjunto de aprendizaje [37]. El procedimiento para la conversión de los vectores de origen es el uso de transformadas lineales [38].

Sean $\boldsymbol{X}$ y $\boldsymbol{Y}$ vectores espectrales reales (sus dimensiones pueden ser diferentes), y sea T una función con dominio en $\boldsymbol{X}$ y rango en $\boldsymbol{Y}$, ($T: \boldsymbol{X} \rightarrow \boldsymbol{Y}$). Se dice que T es una transformación lineal si

a) Para todo $X_1, X_2 \in \boldsymbol{X}, T(X_1 + X_2) = T(X_1) + T(X_2)$ (la función T conserva la suma de vectores).
b) Para todo $X \in \boldsymbol{X}, r \in \mathbb{R}$, $T(rX) = rT(X)$ (la función T conserva la multiplicación escalar).

Si $\boldsymbol{X}$ y $\boldsymbol{Y}$ son vectores espectrales complejos, la definición es la misma excepto en que el escalar es complejo, $r \in \mathbb{C}$. Si $\boldsymbol{X} = \boldsymbol{Y}$ entonces T puede llamarse un operador lineal [39].

CAPÍTULO 3: *DISEÑO E IMPLEMENTACIÓN*

El presente capítulo se presenta el diseño e implementación de un sistema de *Voice Morphing*. Como modelo para el desarrollo de este trabajo de grado se toma como referencia, en el transcurso de las fases de diseño e implementación, el modelo en V descrito en la sección 3.1, para cuya implementación se definen diferentes etapas. La sección 3.2 presenta la definición de los requerimientos del sistema, la sección 3.3 define el diseño funcional y la sección 3.4 el diseño técnico y de componentes.

3.1. MODELO EN V

El modelo en V fue propuesto por Alan Davis a comienzos de los años 90. Se basa en el modelo en cascada puro, pero con la incorporación de actividades de pruebas más efectivas y productivas gracias a la introducción de validaciones durante el desarrollo del proyecto [40]. Debido a que en el modelo tradicional, al realizar las pruebas al finalizar el proyecto, se encuentran defectos o fallas de forma tardía, el modelo en V propone la implementación de las pruebas desde las etapas tempranas del ciclo de vida del proyecto y en paralelo con las actividades de desarrollo [41].

La “V” del nombre del modelo corresponde a la forma como el modelo relaciona las fases de desarrollo con las fases de control de la calidad correspondientes. El brazo izquierdo incluye las tareas de diseño y desarrollo del sistema, y el brazo derecho las medidas de control de calidad de cada etapa. En la unión entre los dos brazos, se encuentra la implementación del producto Figura 3.1 [42].

Teniendo en cuenta las características mencionadas, se hará una adaptación del modelo en V para el desarrollo del presente trabajo.

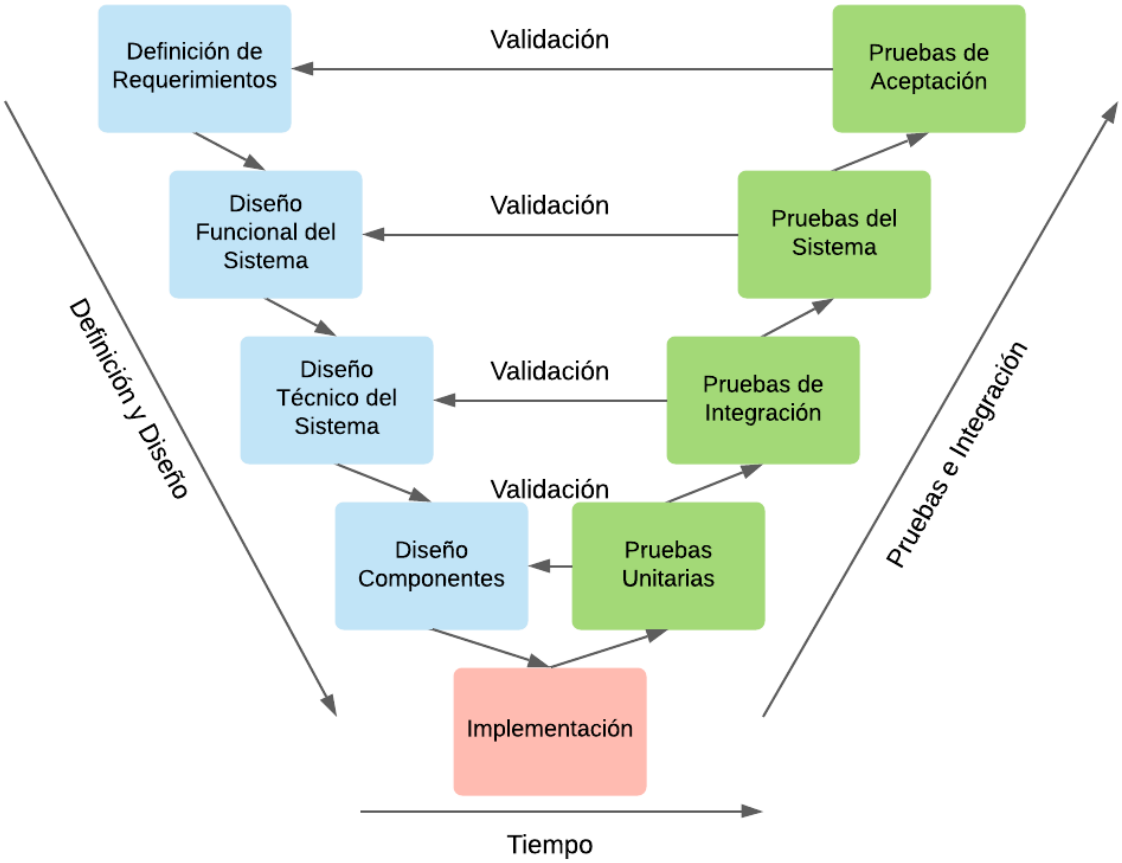


Figura 3.1. Diagrama del modelo en V.
Por los autores.

3.2. DEFINICIÓN DE REQUERIMIENTOS

Para el sistema de *Voice Morphing* se definen dos fases, una fase de entrenamiento y una fase de transformación. Durante el entrenamiento, el sistema analiza y extrae las características de las señales voz del hablante origen y del hablante objetivo, utilizando muestras de voz de entrenamiento previamente almacenadas. En esta etapa, además, se hace el entrenamiento del modelo de conversión basado en la función de transformación.

En la fase de transformación el sistema realiza el mapeo de las características origen a las características objetivo, en función de los datos de entrenamiento de ambos hablantes. El sistema permite modificar algorítmicamente la voz de un hablante origen, para que suene como si hubiese sido pronunciada por un hablante objetivo, conservando el mensaje con una alta precisión y calidad.

Para la evaluación de la calidad de la voz resultante, el sistema cuenta con un algoritmo que le permite analizar las características de la voz obtenida; valorando, de forma subjetiva y objetiva, su naturalidad y proximidad a la voz objetivo.

3.3. DISEÑO FUNCIONAL DEL SISTEMA

Un sistema de *Voice Morphing* de manera general está compuesto por 2 fases principales: Una fase de entrenamiento, en la que se toman las señales de voz origen y objetivo para aplicarles un proceso de análisis y caracterización mediante el cual se estiman las características espectrales de cada una, y en donde se lleva a cabo el entrenamiento del modelo de conversión. Y una fase de transformación, en la que se convierten las tramas de voz del hablante origen en las del hablante objetivo. Dentro de esta última se destaca la etapa de síntesis, en la que las características transformadas del habla se convierten en una señal de voz audible, buscando que sea lo más parecida posible a la señal de voz objetivo. Adicionalmente a esas fases principales, en este trabajo se agrega una fase de evaluación, en la cual se estudia la naturalidad de la voz resultante.

En la fase de entrenamiento, los datos de entrada son las señales de voz origen y objetivo, las cuales pasan por una etapa de preprocesamiento, donde son segmentadas en pequeñas tramas. A dichas tramas se les calcula la frecuencia fundamental y los coeficientes MFCC, características necesarias para la representación de la envolvente espectral de la señal de voz. Para la definición del modelo de conversión, se aplica GMM con el objetivo de estimar y establecer la correlación entre las características espectrales del hablante origen y el hablante objetivo.

En la fase de transformación se aplica una transformación lineal para convertir la envolvente espectral del hablante origen en la del hablante objetivo. Dentro de esta fase se realiza la síntesis de la señal, mediante la cual se toman las características transformadas y se les aplica un proceso de reconstrucción, para conformar una señal de voz audible, la cual corresponde a la señal resultante del sistema.

Finalmente, en la fase de evaluación, se analiza la señal resultante mediante métodos objetivos y subjetivos que permiten estudiar su naturalidad.

En la Figura 3.2 se presenta un diagrama de bloques de la arquitectura del sistema de *Voice Morphing.*

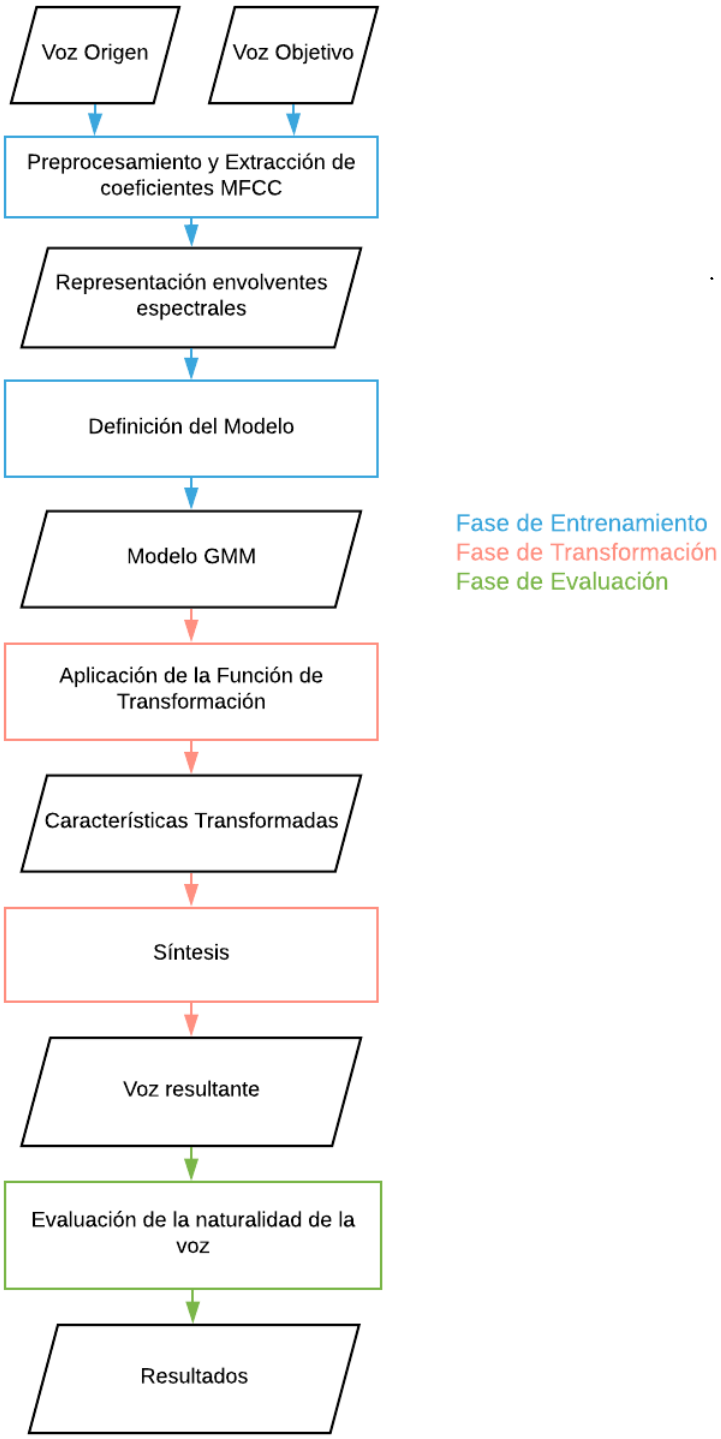


Figura 3. 2. Diagrama de bloques del sistema de *Voice Morphing*.
Por los autores.

3.4. DISEÑO TÉCNICO Y DE COMPONENTES

3.4.1. Bases de datos

El sistema recibe bases de datos de voz tanto del hablante de origen como del hablante objetivo. Para este trabajo se utilizan las bases de datos CMU ARCTIC [43], construidas en el Instituto de Tecnologías del Lenguaje de la Universidad Carnegie Mellon, con el propósito de la investigación de síntesis de voz. Existen 7 conjuntos principales de grabaciones (5 masculinas, 2 femeninas) y varias bases de datos auxiliares. Cada una de las bases de datos corresponde a un hablante y consta de casi 1150 expresiones en inglés equilibradas fonéticamente. Estas expresiones son las mismas para cada uno de los locutores. Dichas bases de datos se distribuyen como *software* libre, sin restricción de uso comercial o no comercial. Las señales de voz fueron grabadas a 16 KHz y se encuentran en formato WAV[6]. De las bases de datos disponibles, para este trabajo de grado, se toman muestras de señales de voz de hablantes masculinos pronunciando las mismas frases en inglés.

3.4.2. Fase de entrenamiento

Dentro de la fase de entrenamiento el primer paso consiste en realizar el preprocesamiento de los datos, con el fin de que se pueda realizar correctamente la posterior estimación de la envolvente espectral. El preprocesamiento de los datos o de las señales de voz está compuesto por un proceso segmentación y cálculo de la frecuencia fundamental.

Una señal de voz es de naturaleza dinámica y cambia con el tiempo, sin embargo, para el análisis del habla se considera que las señales de voz son lo suficientemente estacionarias en segmentos de 5 a 100 ms. La señal de voz se descompone en una serie de segmentos cortos denominadas tramas. Por lo anterior, para el análisis del presente trabajo se define una trama de 5 ms [44].

MFCC

La extracción de las características de las señales de voz de origen y objetivo es una de las tareas más importante para el rendimiento del sistema de *Voice Morphing*. MFCC se fundamenta en las percepciones auditivas humanas lo que

[6] Formato de audio digital con o sin compresión de datos que cuando se utiliza con formato LPCM (sin compresión), presenta la más alta calidad y es adecuado para uso profesional.

resulta en una representación compacta y robusta de la envolvente espectral de la señal de voz [45]. Teniendo en cuenta lo anterior, en el presente trabajo las envolventes espectrales origen y objetivo se parametrizan con coeficientes MFCC, variando entre 12 y 24 coeficientes por trama para cada hablante. Esto con el fin de verificar si al utilizar un mayor número de coeficientes MFCC se logra una representación significativamente mejor de las envolventes espectrales, teniendo en cuenta que esto incrementa los recursos computacionales.

En la Figura 3.3 se muestra el diagrama de bloques de la fase de entrenamiento del sistema de *Voice Morphing*.

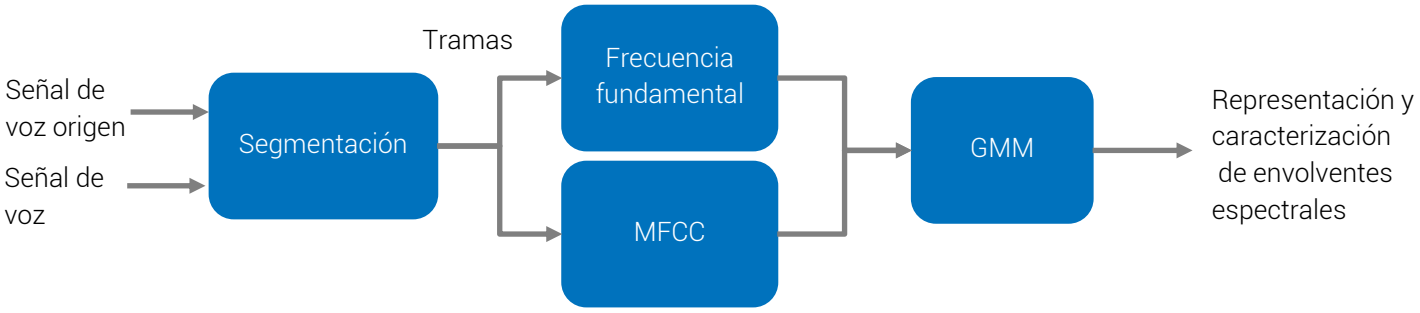


Figura 3.3. Diagrama de bloques fase de entrenamiento del sistema de *Voice Morphing*.
Por los autores.

El resultado de la fase de entrenamiento es la caracterización de las señales de voz de los dos hablantes por medio de sus envolventes espectrales, no obstante, los resultados varían según el número de coeficientes considerados. En la Figura 3.4 y la Figura 3.5 se puede observar la representación espectral de las envolventes de dos hablantes, con 12 y 24 coeficientes MFCC respectivamente.

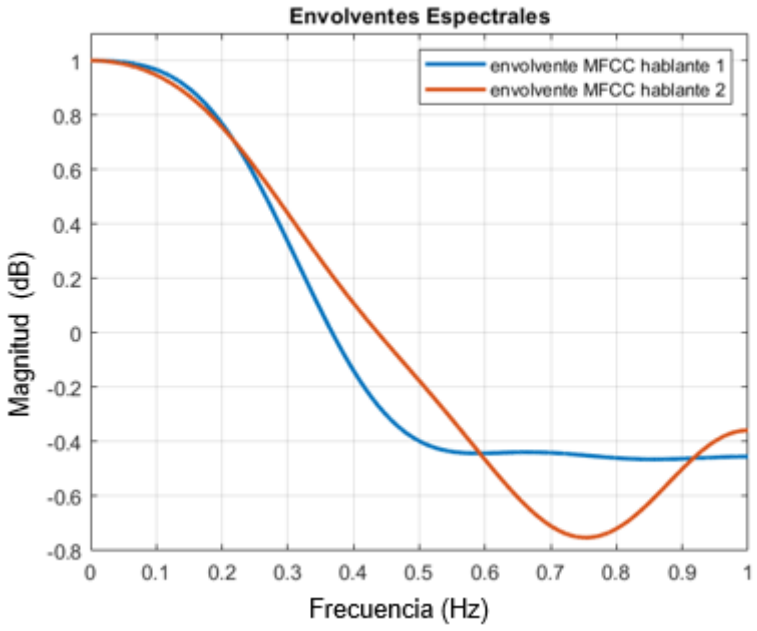


Figura 3.4. Envolventes espectrales con 12 coeficientes MFCC.
Por los autores.

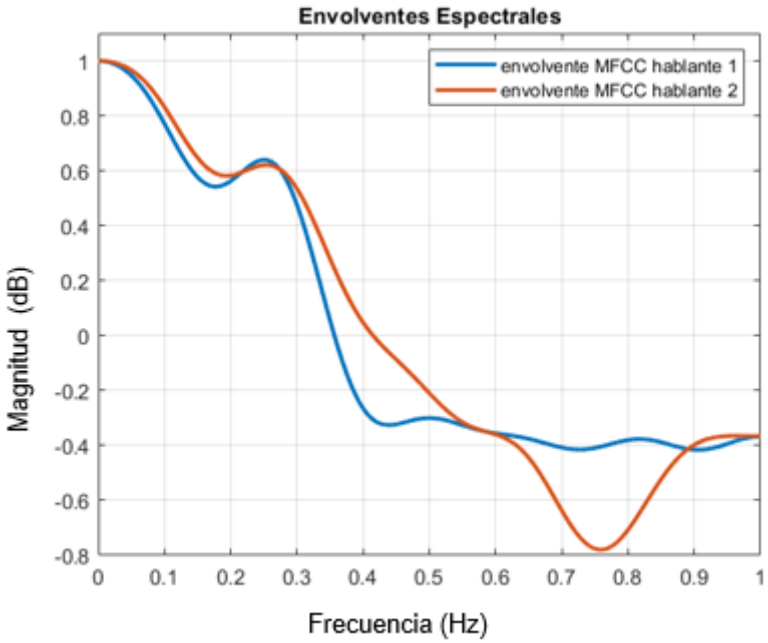


Figura 3.5. Envolventes espectrales con 24 coeficientes MFCC.
Por los autores.

Definición de Modelo

En el sistema de *Voice Morphing* del presente trabajo la función de transformación, para el mapeo de las características origen y objetivo, está fundamentada en el GMM [29]. En este enfoque de transformación de voz, la relación entre los dos conjuntos de envolventes espectrales previamente representadas por medio de los coeficientes MFCC, que corresponden a los hablantes origen y objetivo, se representa utilizando un modelo de mezcla gaussiano (ver Figura 3.6) [46].

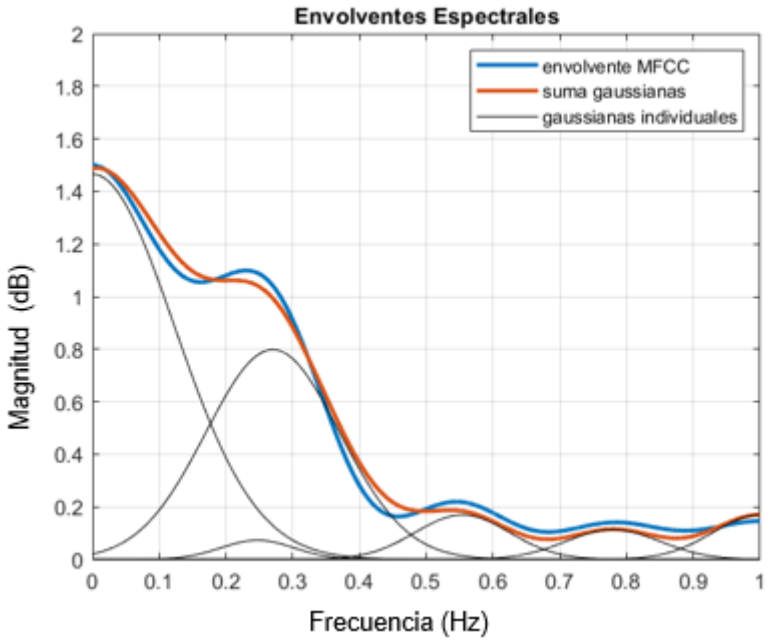


Figura 3.6. Representación de la envolvente espectral a través de GMM.
Por los autores.

Para modelar los datos usando un modelo de mezcla gaussiano y encontrar una transformación lineal se utiliza la densidad conjunta entre hablantes [29].

Para modelar la densidad conjunta de las características de origen y objetivo con un GMM, los vectores origen $\boldsymbol{X}_n$ se concatenan con los vectores de destino $\boldsymbol{Y}_n$ correspondientes

$$\boldsymbol{Z}_n = [\boldsymbol{X}_n^T, \boldsymbol{Y}_n^T]^T \tag{3.1}$$

y $\boldsymbol{Z}_n$ está modelado como

$$p(\boldsymbol{Z}_n) = \sum_{m=1}^{M} \alpha_m \, N\left(\boldsymbol{Z}_n; \mu_m^{(z)}, \Sigma_m^{(z)}\right) \tag{3.2}$$

donde

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix} \tag{3.3}$$

$$\Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix} \tag{3.4}$$

$\mu_m^{(x)}$ y $\mu_m^{(y)}$ denotan la media de las entradas de origen y objetivo del vector aumentado, respectivamente y los superíndices de las matrices de covarianza denotan sus respectivas covarianzas y covarianzas cruzadas.

En los sistemas basados en GMM, existe una configuración fundamental en la forma de la matriz de covarianza que se utiliza. Las matrices de covarianzas completa, diagonal y diagonal en bloque son algunas de las configuraciones más utilizadas. La matriz de covarianza diagonal implica que los elementos del vector de características son independientes y por tanto se transforman independientemente unos de otros; mientras que la matriz de covarianza completa y diagonal en bloque puede modelar explícitamente todas o algunas de las correlaciones respectivamente [47]. En el presente trabajo se descarta la matriz de covarianza diagonal puesto que, teniendo en cuenta que los hablantes origen y objetivo pronuncian la misma frase, entre las dos envolventes espectrales existen ciertas correlaciones. Por tanto, se utiliza la matriz de covarianza completa y la matriz de covarianza diagonal en bloque con el fin de observar con cuál se obtiene una mejor función de mapeo, teniendo en cuenta que el uso de una matriz de covarianza completa incrementa notablemente el número de parámetros a transformar y a su vez el costo computacional.

Otra variable importante en la configuración de sistemas GMM es el número de componentes gaussianos a utilizar para representar las envolventes espectrales de los hablantes origen y objetivo. Utilizar pocas componentes puede generar una representación muy poco aproximada de las envolventes, mientras que un número muy elevado puede resultar en un sobreajuste, que a su vez produce una mayor distorsión del audio transformado. Teniendo en cuenta esto, para el presente trabajo se utilizan configuraciones con 8, 16 y 32 componentes gaussianos [31], [48], [49], puesto que se quiere observar el efecto que tiene cada una sobre la calidad de la voz resultante.

3.4.3. Fase de transformación

Alineamiento

En el sistema de *Voice Morphing* las características del hablante origen y objetivo se alinean a nivel de cuadro para obtener correspondencia entre los diferentes sonidos del habla. Se utiliza la Deformación de Tiempo Dinámica (DTW, *Dynamic Time Warping)*, una de las técnicas más utilizadas para alinear dos oraciones a nivel de trama. Con la aplicación de DTW en el sistema, cada trama de origen se empareja idealmente con su correspondiente trama objetivo [29].

Aplicación de la Función de Transformación

En la fase de transformación la identidad del hablante origen debe convertirse con la mayor fidelidad realizable para la obtención de una voz de alta calidad.

Para la transformación de características, una función de transformación mapea el vector de características origen $\mathbf{X}$ en el vector de características objetivo $\mathbf{\Psi}$ a través del modelo de conversión definido.

En la transformación, el objetivo mapeado $\hat{\boldsymbol{y}}_n$ es formado a partir del vector de origen $\boldsymbol{X}_n$ así:

$$\hat{\boldsymbol{y}} = \sum_{m=1}^{M} \omega_{m,n} \left[\mu_m^{(y)} + \Sigma_m^{(yx)} \left(\Sigma_m^{(xx)} \right)^{-1} \left(\boldsymbol{X}_n - \mu_m^{(x)} \right) \right] \tag{3.5}$$

donde $\omega_{m,n}$ es la probabilidad a posteriori de que el m-ésimo gaussiano haya producido la n-ésima observación (Ecuación 3.6) [50].

$$\omega_{m,n=} \frac{\alpha_m N(\mathrm{x}_n; \mu_m^{(x)}, \Sigma_m^{(x)})}{\sum_{j=1}^{M} \alpha_j \, N(\mathrm{x}_n; \mu_j^{(x)}, \Sigma_j^{(x)})} \tag{3.6}$$

Síntesis

La síntesis es el proceso a través del cual las envolventes y demás características espectrales transformadas son convertidas de nuevo en una señal de voz audible que debe ser lo más parecida posible a la señal de voz del hablante objetivo. La reconstrucción puede entenderse como una función inversa del análisis del habla, que trabaja sobre las características modificadas y produce una señal de voz audible (ver Figura 3.7) [46].

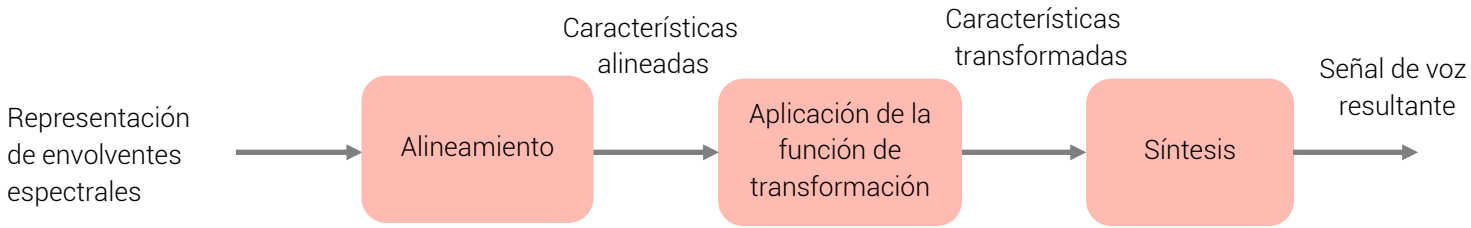


Figura 3.7. Diagrama de Bloques Fase de Transformación del Sistema de *Voice Morphing*.
Por los autores.

Implementación del Sistema de *Voice Morphing*

Para la implementación del sistema de *Voice Morphing* se decide utilizar el lenguaje de programación Python, ya que, a diferencia de MATLAB, este lenguaje cuenta con una función para calcular directamente las envolventes espectrales a partir de los MFCC. Adicionalmente, dentro de su documentación se encuentra un proyecto

Sprocket, un *Software* de *voice Conversion* de código abierto, el cual constituye la base del desarrollo de este trabajo de grado. El proyecto *Sprocket* fue desarrollado por Kazuhiro Kobayashi y Tomoki Toda en la Universidad de Nagoya de Japón.

Sprocket es un software de código abierto que transforma la identidad de un hablante de origen en la de un hablante objetivo utilizando los métodos de *Voice Morphing* basados en GMM con un conjunto de datos paralelos. Su licencia basada en la licencia MIT permite que sus características se puedan utilizar libremente tanto para propósitos de investigación como para fines industriales.

Instalación

La Figura 3.8. muestra la estructura de directorios de sprocket. Python3 se adopta como el lenguaje de programación principal. Inicialmente, es necesario instalar las librerías dependientes mediante el comando *pip*. Luego, ejecutando el comando `python3 setup.py install` en una terminal, las librerías que usa sprocket se instalan en el entorno Python3.

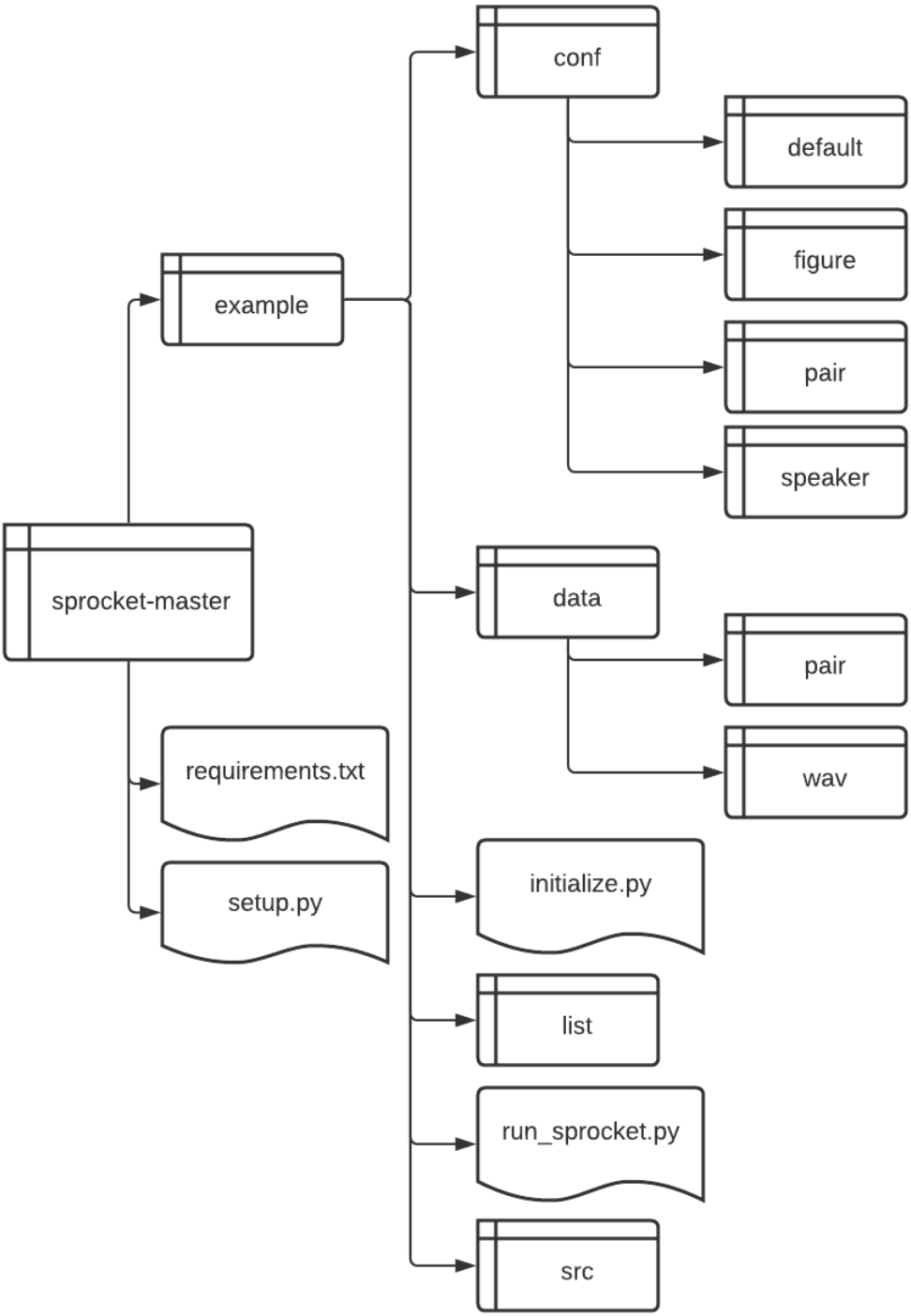


Figura 3. 8. Directorio Proyecto Sprocket.
Por los autores.

Preparación del Conjunto de Datos

Se asume que el directorio de trabajo está configurado en `example/`.

Es necesario preparar un conjunto de datos paralelos que consten de las mismas oraciones pronunciadas por los diferentes hablantes de origen y destino.

El formato de archivo admitido de las señales de voz es 16000 Hz, 22050 Hz, 44100 Hz, o 48000 Hz para la frecuencia de muestreo. Los archivos de audio de los hablantes de origen y destino deben almacenarse en `data/wav` (e.g., `data/wav/hablanteA/*.wav`, para el hablante A).

Inicialización

Para generar los archivos de lista y los archivos de configuración dependientes del hablante y dependientes del par, para su uso en los procesos de entrenamiento y conversión, se ejecuta el archivo `initialize.py`. `initialize.py` toma tres argumentos. El primer argumento corresponde al hablante origen (e.g., hablanteA), el segundo argumento corresponde al hablante objetivo (e.g., hablanteB) y el tercer argumento es para la frecuencia de muestreo del formato (e.g., 16000 Hz).

Las listas muestran las rutas de los archivos de audio (e.g., hablanteA_train.list para el entrenamiento y hablanteA_eval.list para la evaluación) se generan en el directorio `list`. Los archivos YAML dependientes de cada hablante (e.g., hablanteA.yml) que muestran el formato de los archivos y los parámetros para la extracción de características acústicas y el archivo YAML dependiente del par (e.g., hablanteA-hablanteB.yml) que muestra los parámetros utilizados en el modelado GMM se generan en los directorios `conf/speaker` y `conf/par`, respectivamente.

Modificación de las Listas

Es necesario modificar las listas generadas automáticamente para seleccionar los enunciados de entrenamiento y evaluación. Las listas de entrenamiento se utilizan para el cálculo de las estadísticas dependientes del hablante y el modelo GMM. Las listas de evaluación definen los archivos de audio con los cuales se evalúa el modelo. Debido a que los archivos de audio utilizados para el entrenamiento y la evaluación deben ser independientes se hace necesario organizar manualmente las listas de entrenamiento y evaluación. Se debe tener en cuenta que el número de archivos de audio debe ser el mismo para los hablantes de origen y de destino.

Configuración de los Parámetros Dependientes del Par de Hablantes

El archivo YAML de cada uno de los hablantes se genera en el directorio `conf/speaker` mediante la ejecución de `initialize.py`. Aquí se modifican los parámetros para la caracterización de los hablantes como por ejemplo el número de coeficientes MFCC.

Configuración de los Parámetros Dependientes del Par de Hablantes

El archivo YAML dependiente del par de hablantes se genera en el directorio `conf/pair` mediante la ejecución de `initialize.py`. Aquí se modifican los parámetros para la construcción del modelo GMM, como por ejemplo el número de componentes gaussianos y el tipo de matriz de covarianza. Debido a que el número de componentes del GMM afecta en gran medida la calidad de la conversión, debe establecerse cuidadosamente de acuerdo con el número de enunciados de entrenamiento (ver Figura 3.9).

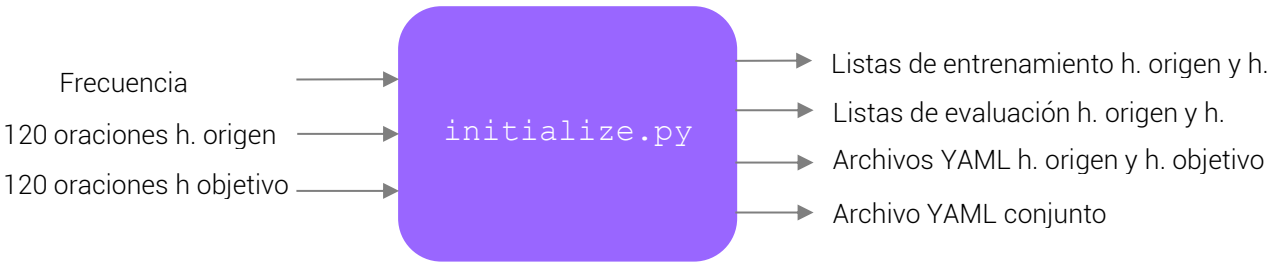


Figura 3. 9. Entradas y salidas de la etapa de inicialización.
Por los autores.

Procesos de Entrenamiento y Transformación

Una vez configurados los parámetros descritos anteriormente en los archivos YAML se procede al entrenamiento y construcción del modelo. El script `run_sprocket.py` toma los dos argumentos correspondientes al hablante origen y objetivo y 5 opciones correspondientes a las etapas descritas a continuación, (e.g., `python3 run_sprocket.py -1 -2 -3 -4 -5 hablanteA hablanteB`) (ver Figura 3.10).

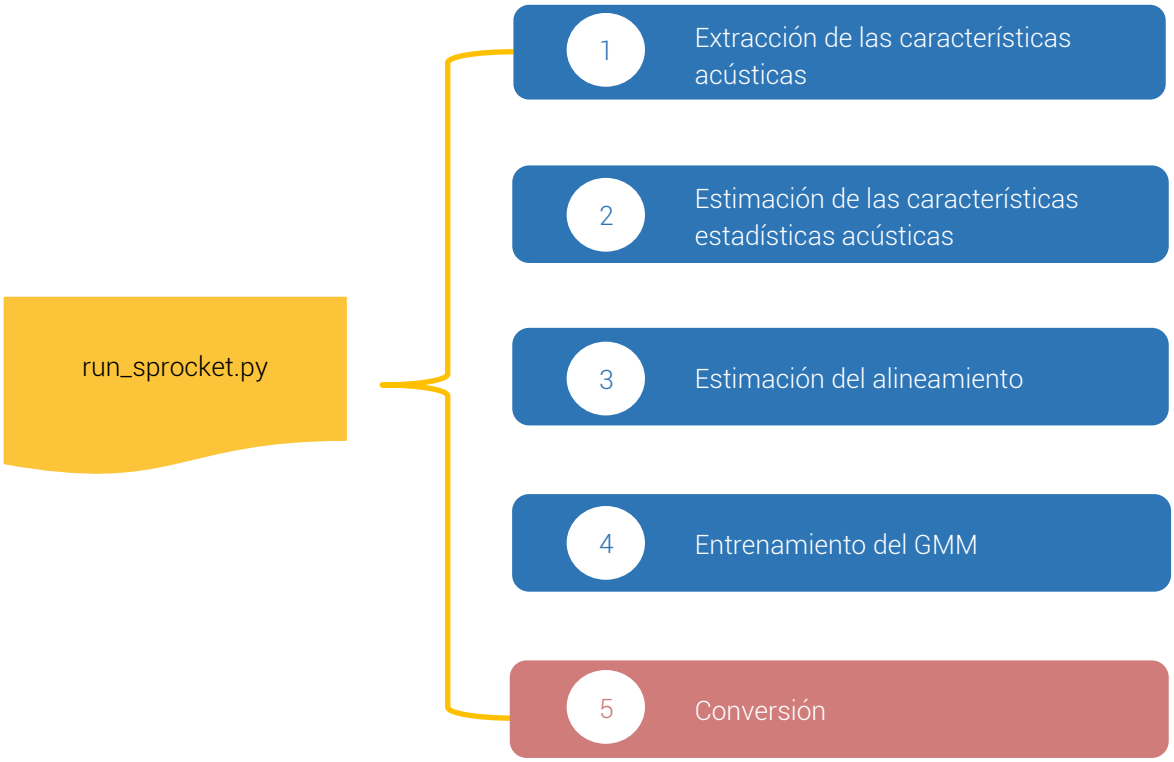


Figura 3. 10. Diagrama de Flujo del Archivo run_sprocket.py.
Por los autores.

1. En `example/src/yml.py` se define los parámetros que caracterizan a los hablantes. Lee los audios de entrenamiento y prueba.

   En `example/src/initialize_speaker.py` se utiliza a `sprocket/speech/feature_extractor.py` para estimar f_0, la secuencia de la envolvente espectral `spc` y la secuencia de aperiodicidad `ap`.

   Para esto se hace uso de:

   - `Pyworld.harvest:` diseñado por M. Morise, estima f_0 a partir de los múltiples filtros pasa banda y secciones de una señal de voz. Se compara la estimación entre secciones, dado que, el perfil de f_0 no cambia drásticamente con el tiempo. El perfil de f_0 (f_0 *contour*) muestra los cambios de f_0 en el tiempo y tiene una gran importancia en la inteligibilidad de la voz [51].

   - `Pyworld.cheaptrick:` también es propuesto por M. Morise y utiliza enventanado para, con la ayuda de la FFT y f_0 calcular una versión suavizada de la envolvente espectral [52].

- `Pyworld.d4c:` algoritmo, propuesto por M. Morise, que busca mejorar la calidad de las señales de voz procesadas, para lo cual realiza una representación estadística temporalmente dependiente. Tener en cuenta la aperiodicidad y la representación estadística (MLE) ayuda a mejorar la calidad del habla y la diferenciación de los hablantes [53].

En `analyzer.py` la envolvente de `spc` tiene que ser la misma que la `ap`.

Se convierte la envolvente espectral en mel-cepstrum a través de sp2mc, así:

a) Envolvente espectral enventanada.
b) Mapeo a la escala de mel utilizando los filtros triangulares con traslape.
c) Logaritmo de las frecuencias mel.
d) DCT de lo anterior.

A la entrada de `sp2mc` se ingresa la envolvente espectral normalizada. La normalización se hace en `sprocket/speech/parameterizer.py` con `spc2npow` y se invoca en `FeatureExtractor` con `npow`.

En `sprocket/speech/synthesizer.py` se crea una forma de onda a partir de f_0, mcep y la aperiodicidad. Se calcula nuevamente la secuencia de la envolvente espectral a partir de los mcep, para esto usa la función `pysptk.mc2sp`. Luego con `pyworld.synthesize` genera la forma de onda.

En `extract_features.py` se guardan en los HDF5 los siguientes datos: f_0, mcep, npow, codeap. La forma de onda generada con `synthesizer.py` se guarda como un `.wav`

2. En `estimate_feature_statistics.py` se tiene:

- `sprocket/model/gv.py` estima una varianza global. `gv.estimate` se invoca en la línea 47 de `estimate_feature_statistics.py` y esta información se agrega al archivo HDF5.

- `sprocket/model/f0statistics.py`, aquí, a partir de la lista de valores de $f_0$convertidos a escala logarítmica, se calcula la media y la desviación estándar. Se invoca en la línea 40 de `estimate_feature_statistics` y se agrega al archivo HDF5 (ver Figura 3.11).

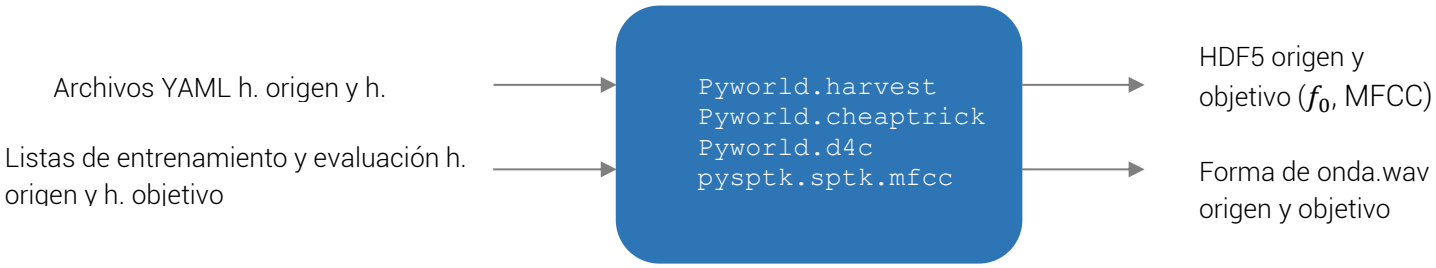


Figura 3. 11. Entradas y salidas de las etapas 1 y 2 de run_sproket.py.
Por los autores.

Dado que para la transformación se utiliza la conjunta de los dos hablantes (ver ecuación 3.1).

3. En `example/src/estimate_twf_and_jnt.py` se realiza la estimación del vector conjunto de características de la pareja de hablantes utilizando GMM:

   I. Se lee los archivos de cada hablante generados en la opción 1.

   II. Se alinean los vectores de `org_mceps` y `tar_mceps`. Para esto se utilizan los vectores de potencia normalizado de origen y objetivo y la función `extsddata` para diferenciar las tramas de silencio. El alineamiento se hace con DTW con la función `align_data`, también se calcula la distorsión introducida en los coeficientes con la función `melcd`.
   Dentro de `misc` se encuentra la función/método `transform_jnt` que con `numpy.r` concatena las funciones de DTW, twf (ver Figura 3.12).

Figura 3. 12. Entradas y salidas de la etapa de estimación de alineamiento de run_sproket.py.
Por los autores.

4. Se entrena el modelo GMM a partir de `jdata = align_data(oexdata, texdata, twf)` es decir los coeficientes umbralizados, alineados y

concatenados. Para el entrenamiento se define un objeto, `param`, en el cual se especifican: el número de componentes, el tipo de matriz de covarianza y el número de iteraciones que realizará el algoritmo. Además, se debe define si la técnica es:

- Generación de Parámetros de Máxima Verosimilitud (MLPG, *Máximum Likelihood Parameter Generation*).
- Error Cuadrático Medio Mínimo (MMSE, *Minimum mean square error*).

Esto se hace en `GMM {mcep, cvtype}` especificado en el archivo de configuración del par de hablantes: `hablanteA-hablanteB.YML`. Para el ajuste del modelo se utiliza la función `fit`, la cual estima los parámetros del modelo a partir del algoritmo EM (ver Figura 3.13). Estos parámetros se almacenan en el directorio `example/data/pair/hablanteA-hablanteB/modelo`.

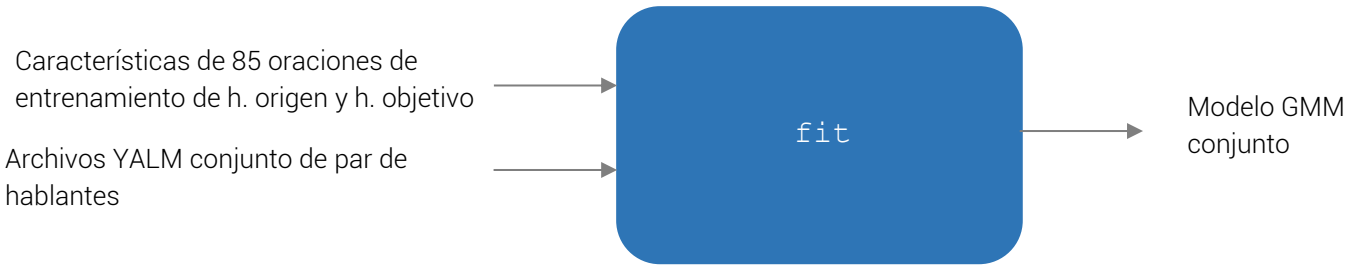


Figura 3. 13. Entradas y salidas de la etapa de entrenamiento del modelo GMM de run_sprocket.py. Por los autores.

Las muestras de voz de origen, descritas en las listas de evaluación del hablante origen, se convierten en muestras de la voz objetivo con la función `convert` (ver Figura 3.14). Las voces convertidas se etiquetan como * VC.wav. Estas voces convertidas se guardan en el directorio example/data/pair/hablanteA-hablanteB/test.

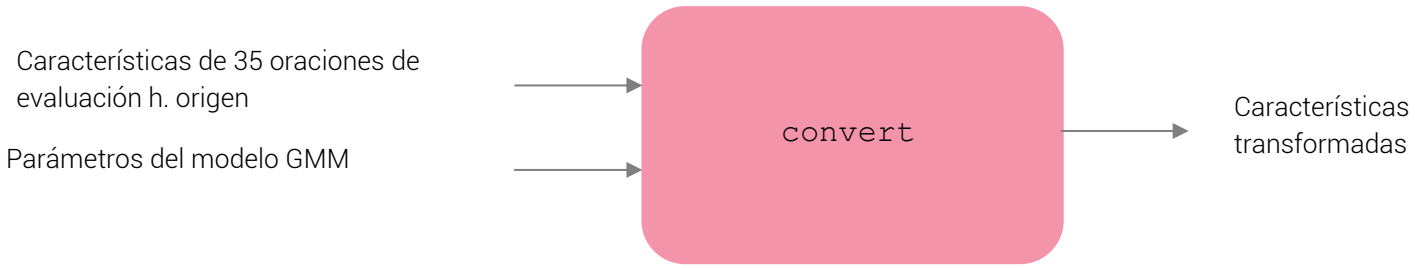


Figura 3. 14. Entradas y salidas de la etapa de conversión de run_sprocket.py. Por los autores.

En la Figura 3.15 se muestra la señal de voz del hablante objetivo versus la señal de voz convertida, en el dominio del tiempo. En la Figura 3.16 se muestra la señal de voz del hablante objetivo versus la señal de voz convertida, en el dominio de la frecuencia. A partir de estos resultados se observa que las diferencias entre los espectros de magnitud y la omisión de la información del espectro de fase provocan que las señales en el dominio del tiempo no sean semejantes, no obstante, el oído humano tiene la capacidad de compensar pequeñas distorsiones de fase e interpretar correctamente la señal de voz convertida.

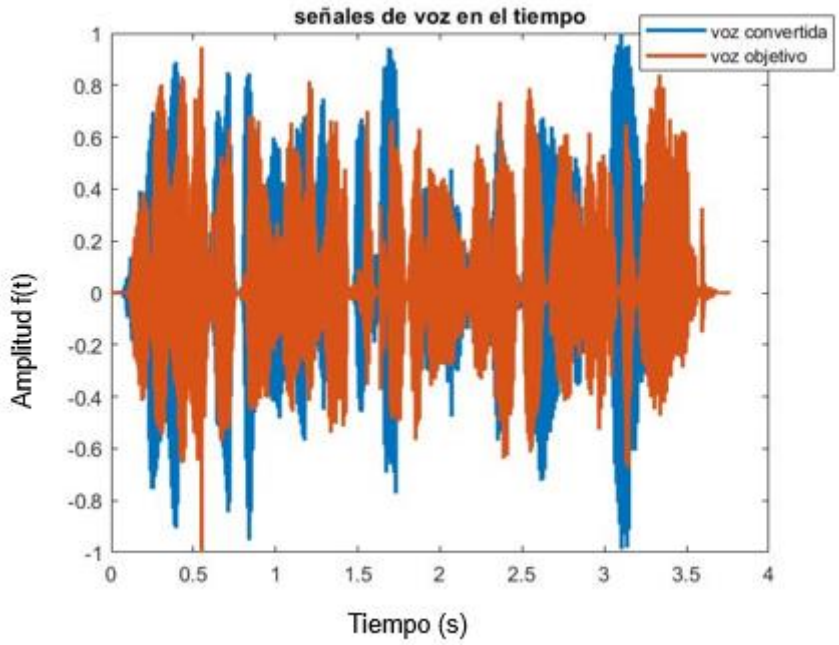


Figura 3. 15. Señal de voz del hablante objetivo vs señal de voz convertida, en el dominio del tiempo. Por los autores.

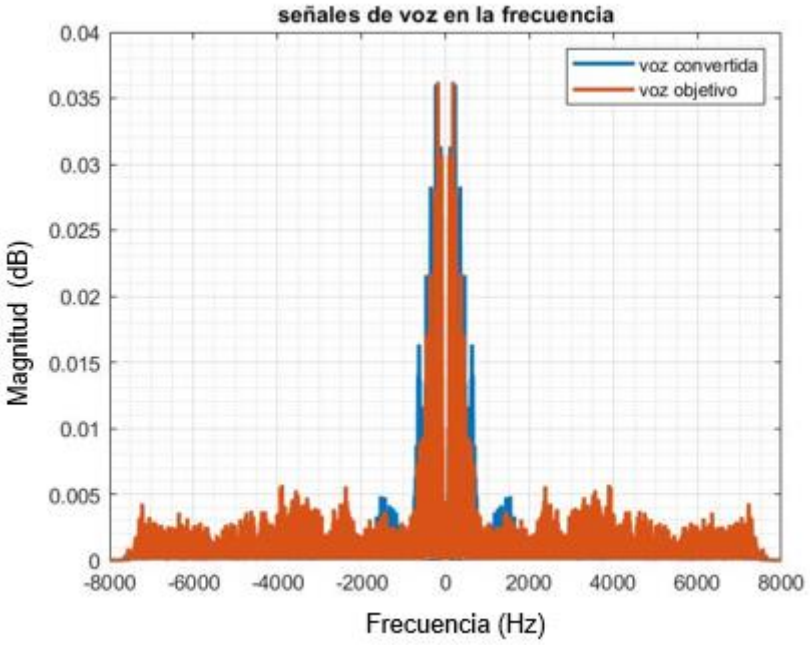


Figura 3. 16. Espectros de la señal de voz del hablante objetivo vs señal de voz convertida.
Por los autores.

La fuente del código del proyecto base, proyecto Sprocket y las bases de datos, utilizados en la implementación del sistema de *Voice Morphing,* se encuentra detallado en el Apéndice B.

3.4.4. Evaluación subjetiva y objetiva

Es necesaria una evaluación efectiva de la calidad de la voz resultante, para validar los algoritmos y técnicas utilizadas en conjunto, en el sistema de *Voice Morphing* implementado. Generalmente, los resultados se presentan en términos de mediciones objetivas y subjetivas [46].

La calidad de una señal de voz es necesariamente una medida subjetiva. A pesar de que la implementación de una evaluación subjetiva es costosa y demanda mucho tiempo y de que, además, existan medidas objetivas, siempre se deben emplear pruebas subjetivas para determinar la calidad final de la voz resultante [29].

Evaluación objetiva

El rendimiento del sistema se estima objetivamente contrastando las características de voz transformadas con las características de voz objetivo. Para una evaluación objetiva de sistemas de síntesis y transformación de voz existen una gran variedad de técnicas y algoritmos. Para el presente trabajo se ha seleccionado la Distorsión Mel-cepstral (MCD, *Mel-cepstral distortion*) para la evaluación de transformación de espectro y el Error Cuadrático Medio (RMSE, *Root Mean Square Error*) para la evaluación de transformación de prosodia [46], debido a las características del sistema de *Voice Morphing* y la efectividad de los algoritmos.

Transformación de espectro

Para la evaluación objetiva de la transformación espectro del sistema, se estiman las diferencias entre el habla transformada y el habla objetivo, comparando sus distancias espectrales. Para garantizar que la voz transformada y la voz de objetivo tienen la misma longitud se utiliza un alineador de cuadros para establecer el mapeo a nivel de cuadros. Luego se calcula la MCD, criterio de error objetivo de percepción, utilizado para medir la diferencia entre dos características espectrales; muy útil en estudios estadísticos de síntesis de voz y transformación de voz.

La MCD se estima entre los coeficientes cepstrales de Mel del habla transformada y el habla objetivo y se calcula como [29]:

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2} \tag{3.7}$$

donde c_d y $\hat{c}_d$ son el d-ésimo MCC del objetivo original y convertido, respectivamente, y D es el orden de los MCC [29].

Una MCD más baja indica un mejor rendimiento del sistema. Debido a que el valor de la MCD no siempre se correlaciona con la percepción humana es necesario complementar la evaluación del sistema de *Voice Morphing* con pruebas subjetivas.

Transformación de la prosodia

La prosodia del habla está determinada por la duración fonética, el contorno de energía y el contorno de tono [46]. Para calcular la proximidad entre los patrones de prosodia del habla transformada y del habla objetivo se hace uso de RMSE, como métrica de evaluación, para medir la dependencia lineal de los contornos de prosodia o de energía.

El RMSE entre el habla transformada y el habla objetivo se define como:

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F0_k^c - F0_k^t)^2} \tag{3.8}$$

donde $\mathrm{F0_k^c}$ y $\mathrm{F0_k^t}$ son las características del habla transformada y objetivo, respectivamente. K es la longitud de la secuencia $\mathrm{F0}$ o el número total de tramas. Un valor de RMSE bajo denota un sistema eficiente de transformación de $\mathrm{F0}$ [46]. En la Figura 3.17 se presenta el diagrama de bloque de la evaluación objetiva del sistema.

Figura 3.17. Diagrama de bloques evaluación objetiva del sistema.

Por los autores.

Evaluación subjetiva

Las métricas de la evaluación subjetiva se utilizan para valorar tanto la naturalidad de la voz resultante como la similitud con la voz objetivo, es decir la identidad.

Existen un gran número de medidas subjetivas para la evaluación de la naturalidad de la señal de voz transformada y su identidad [46]. Para el presente trabajo se han seleccionado la prueba de puntuación de opinión media (MOS, *Mean Opinion Score)* y la prueba ABX teniendo en cuenta el número limitado de oyentes y las características del sistema.

MOS

En esta prueba los oyentes califican una serie de archivos de audio utilizando una escala de deterioro de cinco niveles: (*malo=1, pobre=2, regular=3, bueno=4, excelente=5*). Después de escuchar cada muestra, los oyentes expresan una opinión, fundamentada únicamente en la muestra escuchada recientemente. El promedio de todos los puntajes obtenidos para el habla sintetizada por el sistema representa su puntaje de opinión promedio (MOS) [46].

Una de las principales ventajas del MOS es que aporta eficientemente retroalimentación sobre la naturalidad de una voz sintética según la evaluación de los oyentes. Además, no requiere procedimientos de estandarización y calibración del oyente, estímulos de habla preespecificados, entornos de prueba y otros requisitos de procedimiento rígidos como lo hacen otros tipos de pruebas; lo que hace de esta herramienta flexible para distintos objetivos evaluativos y aplicaciones de voz sintetizada [54].

Prueba ABX

Para la evaluación de la identidad de la voz transformada se implementa una prueba ABX. En esta prueba el oyente escucha tres muestras, A, B y X. A y B incluyen muestras de los hablantes origen y objetivo originales de forma aleatoria y X corresponde a la muestra transformada. Se le pide al participante que seleccione cuál de las muestras, A o B, está más cerca de X. Si X está más cerca del hablante objetivo, el sistema funciona correctamente [29].

En la Figura 3.18 se presenta el diagrama de bloques de la evaluación subjetiva del sistema.

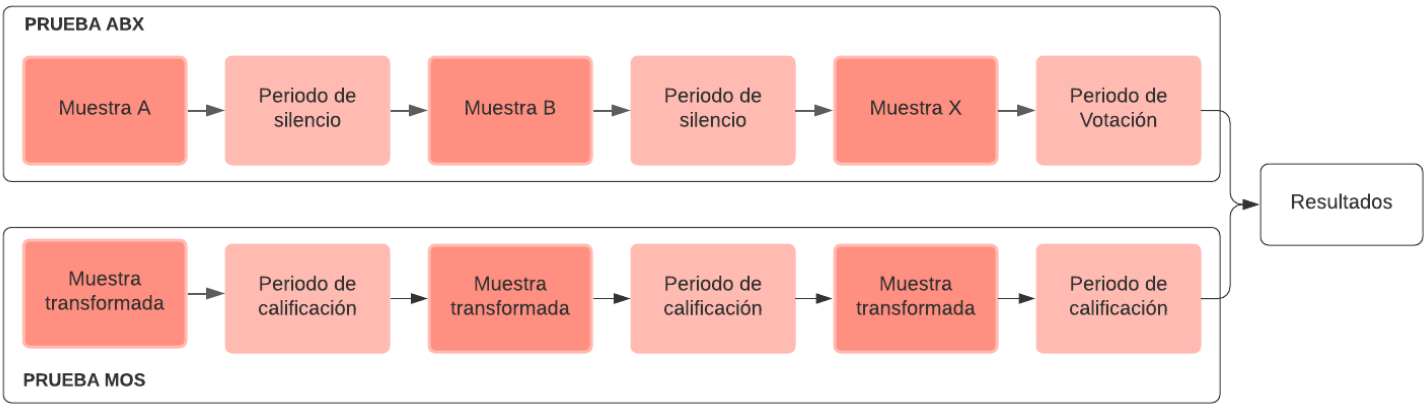


Figura 3. 18. Diagrama de bloques evaluación subjetiva del sistema.
Por los autores.

CAPÍTULO 4: *PRUEBAS Y RESULTADOS*

En este capítulo se exponen los resultados de las pruebas objetivas y subjetivas del sistema de *Voice Morphing* implementado. En la sección 4.1 se presentan algunas especificaciones para tener en cuenta en el desarrollo de las pruebas. En la sección 4.2 se exponen los resultados de las pruebas objetivas y subjetivas, así como el análisis correspondiente.

4.1. PRUEBAS

4.1.1. Condiciones del Sistema de Línea de Base

Para la evaluación del sistema de *Voice Morphing* se utilizan bases de datos correspondientes a 5 hombres, con las cuales se conforman 12 parejas distintas de hablante origen y objetivo. El número de oraciones utilizadas por cada hablante fue de 120, 85 expresiones para el entrenamiento y 35 para la evaluación. La frecuencia de muestreo establecida es 16KHz. En la Tabla 4.1 se muestran las parejas de hablantes establecidas.

Tabla 4. 1. Parejas de hablante origen y hablante objetivo.

Pareja	Hablante Origen	Hablante Objetivo
1	JMK-Hombre Canadiense	BDL- Hombre Estadounidense
2	RMS-Hombre Estadounidense	BDL- Hombre Estadounidense
3	JMK-Hombre Canadiense	RMS-Hombre Estadounidense
4	RMS-Hombre Estadounidense	AWB-Hombre Escocés
5	JMK-Hombre Canadiense	KSP-Hombre Indio
6	BDL- Hombre Estadounidense	JMK-Hombre Canadiense
7	KSP-Hombre Indio	BDL- Hombre Estadounidense
8	AWB-Hombre Escocés	KSP-Hombre Indio
9	KSP-Hombre Indio	RMS-Hombre Estadounidense
10	AWB-Hombre Escocés	JMK-Hombre Canadiense
11	BDL- Hombre Estadounidense	RMS-Hombre Estadounidense
12	BDL- Hombre Estadounidense	AWB-Hombre Escocés

4.1.2. Variación de Parámetros

El desempeño del sistema de Voice Morphing se evalúa de forma objetiva y subjetiva, analizando la voz transformada para distintas configuraciones de tipo de matriz de covarianza, número de componentes gaussianos y número de coeficientes MFCC.

Para conformar los distintos escenarios del sistema la matriz de covarianza se varía entre completa y diagonal en bloque; el número de componentes gaussianos se selecciona entre el conjunto {8,16,32} y el número de coeficientes MFCC se varía

entre 12 y 24. Estas configuraciones se pueden observar en la Tabla 4.2 y se aplican para cada una de las 4 parejas de hablantes.

Tabla 4.2. Configuraciones del sistema evaluadas.

Pareja	ID	Matriz de Covarianza	Número de Componentes Gaussianos	Número de Coeficientes MFCC
Hablante Origen - Hablante Objetivo	1	Completa	32	24
	2	Diagonal en bloque	32	24
	3	Completa	32	12
	4	Diagonal en bloque	32	12
	5	Completa	16	24
	6	Diagonal en bloque	16	24
	7	Completa	16	12
	8	Diagonal en bloque	16	12
	9	Completa	8	24
	10	Diagonal en bloque	8	24
	11	Completa	8	12
	12	Diagonal en bloque	8	12

Los 3 mejores escenarios de acuerdo con la evaluación objetiva se evalúan de forma subjetiva. El número de oyentes de las pruebas subjetivas es de 23. Debido a que las bases de datos utilizadas son en inglés, se eligieron oyentes nativos o con un nivel alto en este idioma. El formulario diseñado para la evaluación subjetiva del sistema se encuentra detallado en el Apéndice C.

4.2. RESULTADOS

4.2.1. Resultados Evaluación Objetiva

En la Tabla 4.3 se muestran los resultados de las medidas de MDC y RSME para cada una de las configuraciones.

Tabla 4.3. Resultados de la prueba objetiva para cada una de las configuraciones y parejas evaluadas.

			Pareja											
			1	2	3	4	5	6	7	8	9	10	11	12
Configuración	1	MDC	4,1509	4,13	4,1401	3,6597	4,7166	3,9733	4,7501	4,3626	4,593	4,2619	3,845	3,4671
		RMSE	0,1352	0,1345	0,1348	0,1192	0,1536	0,1294	0,1547	0,1421	0,1496	0,1388	0,1252	0,1129
	2	MDC	4,3841	4,2739	4,2887	3,7903	5,0965	3,959	4,8319	4,4922	4,657	4,2943	3,9894	3,6096
		RMSE	0,1428	0,1392	0,1397	0,1234	0,166	0,1289	0,1573	0,1463	0,1516	0,1398	0,1299	0,1175
	3	MDC	4,3006	4,2573	4,4732	3,6089	4,9629	4,0931	4,8448	4,5329	4,8975	4,3538	4,1844	3,4625
		RMSE	0,14	0,1386	0,1457	0,1175	0,1616	0,1333	0,1578	0,1476	0,1595	0,1418	0,1363	0,1128
	4	MDC	4,4028	4,3497	4,5532	3,7425	5,1888	4,1367	4,8948	4,5816	4,9457	4,3741	4,2852	3,5617
		RMSE	0,1434	0,1416	0,1483	0,1219	0,169	0,1347	0,1594	0,1492	0,161	0,1424	0,1395	0,116
	5	MDC	4,2424	4,1053	4,1519	3,6865	4,6959	3,9496	4,7983	4,3963	4,6429	4,2794	3,8846	3,4927
		RMSE	0,1381	0,1337	0,1352	0,12	0,1529	0,1286	0,1562	0,1432	0,1512	0,1394	0,1265	0,1137
	6	MDC	4,4011	4,3503	4,3211	3,8861	5,0907	4,0384	4,8914	4,5635	4,7152	4,3931	4,0301	3,6722
		RMSE	0,1433	0,1417	0,1407	0,1265	0,1658	0,1315	0,1593	0,1486	0,1535	0,1431	0,1312	0,1196
	7	MDC	4,3439	4,3085	4,4753	3,6374	4,9473	4,0979	4,8892	4,5874	4,9267	4,28	4,2206	3,4654
		RMSE	0,1415	0,1403	0,1457	0,1184	0,1611	0,1334	0,1592	0,1494	0,1604	0,1394	0,1374	0,1128
	8	MDC	4,4328	4,4069	4,5993	3,7926	5,1581	4,1644	4,9417	4,6063	4,9935	4,4253	4,3132	3,6034
		RMSE	0,1443	0,1435	0,1498	0,1235	0,168	0,1356	0,1609	0,15	0,1626	0,1441	0,1405	0,1173
	9	MDC	4,1992	4,1509	4,1698	3,7003	4,7508	3,9392	4,7845	4,4606	4,6352	4,3547	3,9012	3,4978
		RMSE	0,1367	0,1352	0,1358	0,1205	0,1547	0,1283	0,1558	0,1453	0,1509	0,1418	0,127	0,1139
	10	MDC	4,4468	4,4342	4,3713	3,9971	5,1329	4,1258	5,0308	4,5512	4,808	4,5238	4,0951	3,7823
		RMSE	0,1448	0,1444	0,1423	0,1302	0,1671	0,1343	0,1638	0,1482	0,1566	0,1473	0,1333	0,1232
	11	MDC	4,3212	4,3129	4,4912	3,6745	4,9623	4,1167	4,849	4,5834	4,9402	4,4655	4,2072	3,5124
		RMSE	0,1407	0,1404	0,1462	0,1197	0,1616	0,1341	0,1579	0,1493	0,1609	0,1454	0,137	0,1124
	12	MDC	4,5056	4,4787	4,6049	3,8892	5,0983	4,257	5,0614	4,6702	5,0189	4,5587	4,3753	3,6921
		RMSE	0,1467	0,1458	0,15	0,1266	0,166	0,1386	0,1648	0,1521	0,1634	0,1484	0,1425	0,1202

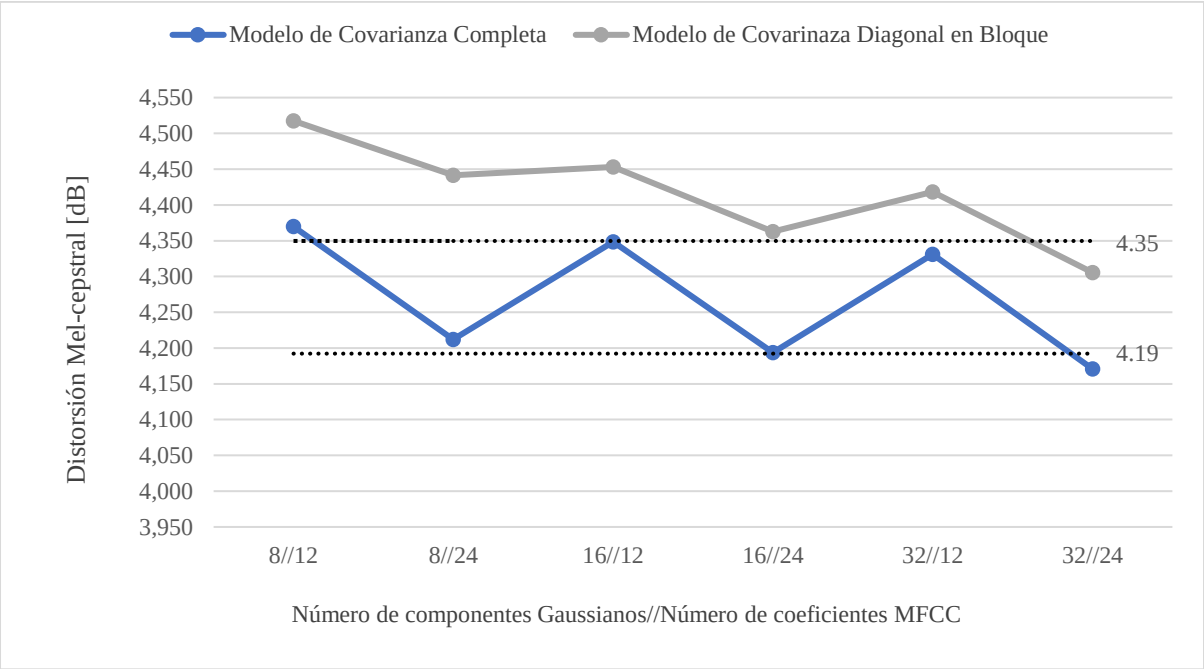


Figura 4.1. Distorsión Mel-cepstral de acuerdo con el tipo de matriz de covarianza del modelo. Por los autores.

En la Figura 4.1 se presenta el promedio de la distorsión Mel-cepstral para cada una de las 12 configuraciones. La línea azul muestra el comportamiento de las configuraciones con matriz de covarianza completa y la línea gris las configuraciones con matriz de covarianza diagonal en bloque.

Se observa que para las configuraciones con matriz de covarianza completa la variación del número de coeficientes MFCC afecta considerablemente los resultados. Por el contrario, la variación de componentes gaussianos no es muy relevante, con un promedio de 4.35 dB para un número de coeficientes igual 12 y 4.19 dB para 24 coeficientes (líneas punteadas).

Para la matriz de covarianza diagonal en bloque se observa que el tanto el número de coeficientes como el número de componentes gaussianos afectan los resultados de la distorsión, este efecto se puede notar principalmente cuando se hace la comparación entre las configuraciones con 8 y 32 componentes.

Desde una perspectiva general, al utilizar una matriz de covarianza completa, se obtienen resultados notablemente mejores que al utilizar matriz de covarianza diagonal en bloque en todas las configuraciones, con un promedio de diferencia de 0.15 dB entre ellas. Se observa además que la distorsión tiende a disminuir en las configuraciones con matriz de covarianza diagonal cuando se aumenta el número de componentes gaussianos. Teniendo en cuenta lo anterior, es posible afirmar que más allá de la representación de las envolventes espectrales de los hablantes, es

más relevante el tipo de matriz de covarianza que se utiliza en el proceso de transformación.

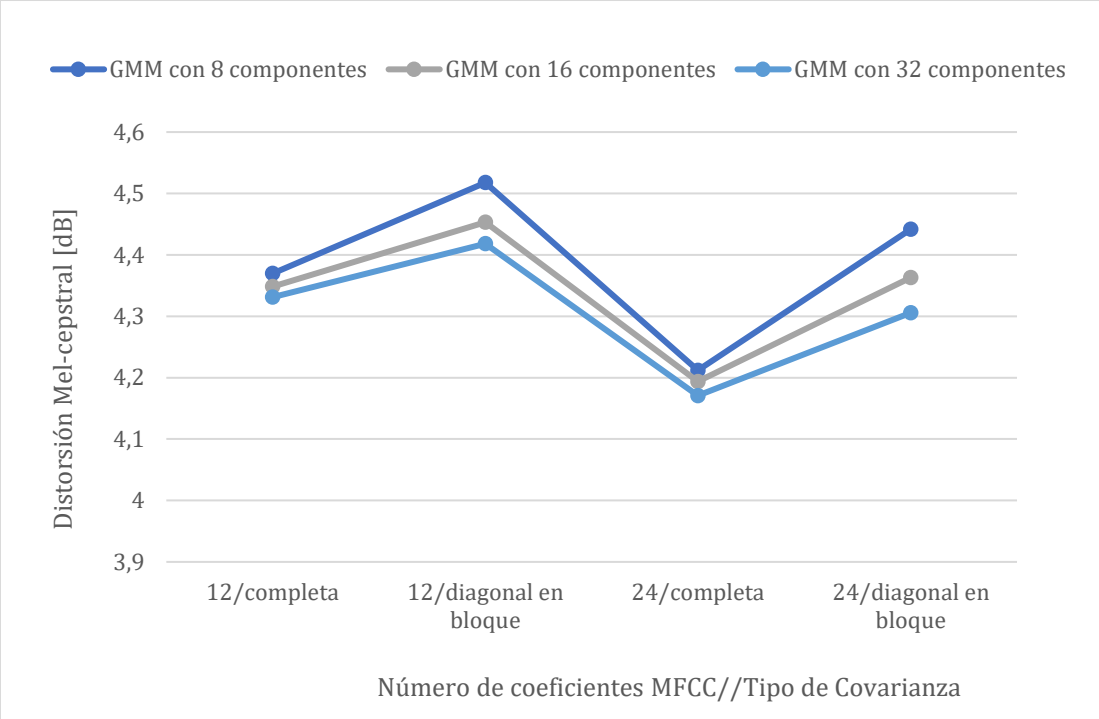


Figura 4.2. Distorsión Mel-cepstral de acuerdo con el número de componentes gaussianos del modelo. Por los autores.

En la Figura 4.2 se muestra el comportamiento de las distintas configuraciones con respecto a la variación del número de componentes gaussianos. Se observa que el número de componentes tiene un mayor impacto en la distorsión cuando la matriz de covarianza es diagonal en bloque con un promedio de diferencia de 0.08 dB entre las configuraciones, mientras que en la matriz de covarianza completa existe una variación menor con un promedio de 0.03 dB.

En general se observa que a un mayor número de componentes existe una tendencia a una disminución de la distorsión en las muestras transformadas, aunque, para una matriz de covarianza completa la variación de este parámetro no representa un cambio considerable en los resultados. A partir de esta observación, es posible afirmar que un mayor número de componentes gaussianos permite una mejor representación de las envolventes espectrales de los hablantes, sin embargo, cuando para la configuración del sistema se utiliza una matriz de covarianza completa se hace menos notable el efecto que tiene el número de coeficientes gaussianos en el resultado final, haciendo del tipo de matriz de covarianza un parámetro crucial.

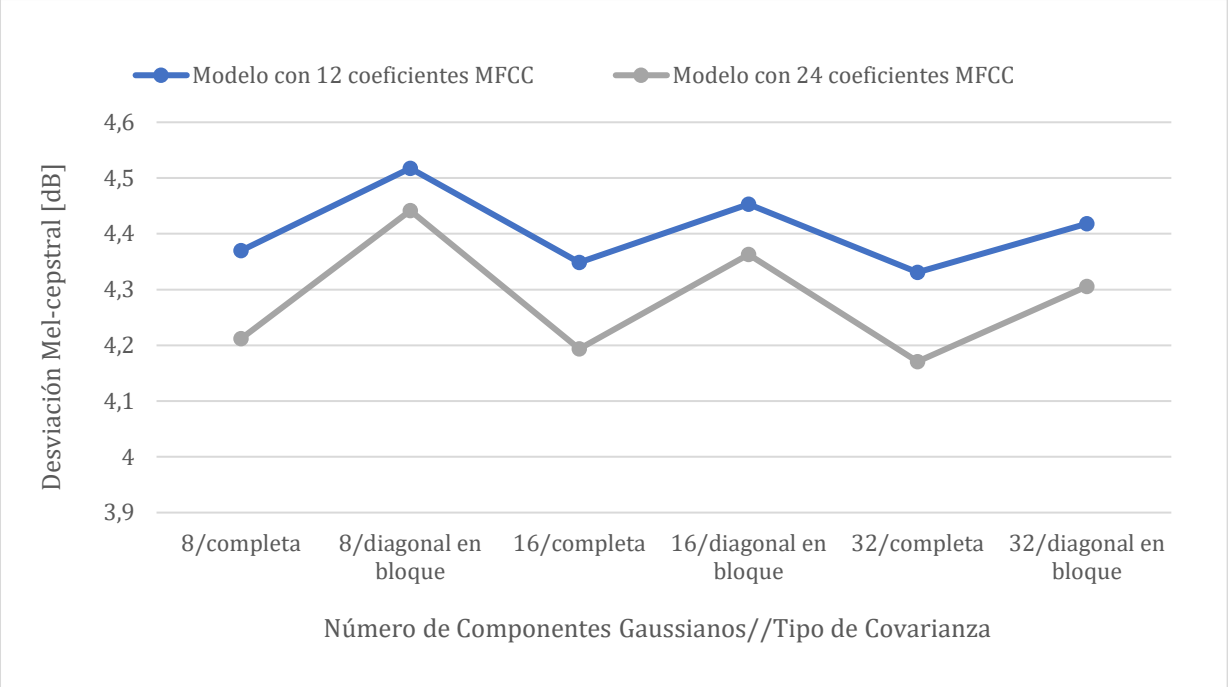


Figura 4.3. Distorsión Mel-cepstral de acuerdo con el número de coeficientes MFCC del modelo.
Por los autores.

En la Figura 4.3 se presenta el promedio de la distorsión Mel-cepstral para cada una de las 12 configuraciones. La línea azul muestra el comportamiento de las configuraciones con número de coeficientes MFCC igual a 12 y la línea gris las configuraciones con un número igual a 24.

Para todas las configuraciones, los valores de distorsión más bajos se obtienen cuando el número de coeficientes MFCC es de 24, esto debido a que al utilizar un mayor número de coeficientes MFCC se obtiene una mejor caracterización de las señales de voz de los dos hablantes.

Se observa que la diferencia de las medidas de la distorsión entre configuraciones es más considerable cuando la matriz de covarianza es completa, en comparación a cuando es diagonal en bloque, con un promedio de variación de 0.16 y 0.09 dB respectivamente. Además, es importante resaltar que, con una matriz de covarianza diagonal en bloque, la variación del número de coeficientes MFCC y el número de componentes gaussianos es posible obtener un rendimiento equivalente a una configuración con matriz de covarianza completa.

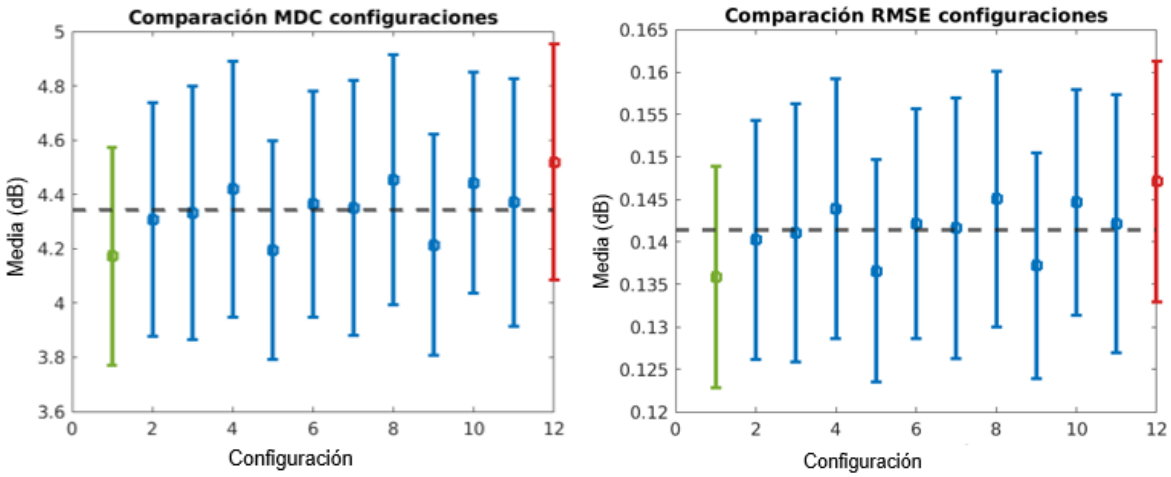


Figura 4.4. Medidas de MCD y RMSE para las 12 configuraciones evaluadas.
Por los autores.

En la Figura 4.4 se presenta el comportamiento de cada una de las 12 configuraciones evaluadas según las medidas de MCD y RMSE. Se resalta en verde la configuración número 1, que obtiene las mejores medidas objetivas. Por otra parte, se resalta en rojo la configuración número 12, que obtiene el rendimiento más bajo según estas medidas.

Se resalta en color verde la configuración con mejor desempeño respecto a las medidas cuantitativas, siendo ésta la configuración con una matriz de covarianza completa, con 24 coeficientes MFCC y con 32 componentes gaussianos. Por otra parte, en color rojo se resalta la configuración con las medidas de distorsión más elevadas, siendo ésta la configuración con matriz de covarianza diagonal en bloque, 12 coeficientes MFCC y con 8 componentes gaussianos.

Se observa, además, que la configuración número 5, correspondiente a la configuración con matriz de covarianza completa, 24 coeficientes y 16 componentes gaussianos, proporciona resultados relevantes, muy cercanos a la mejor configuración. Teniendo en cuenta lo anterior se seleccionan las configuraciones 1, 5 y 12 para el análisis cualitativo.

4.2.2. Resultados Evaluación Subjetiva

Prueba ABX

A continuación, se presentan los resultados obtenidos en la prueba ABX para cada una de las 3 configuraciones. Para el análisis, A corresponde al hablante objetivo y B al hablante origen.

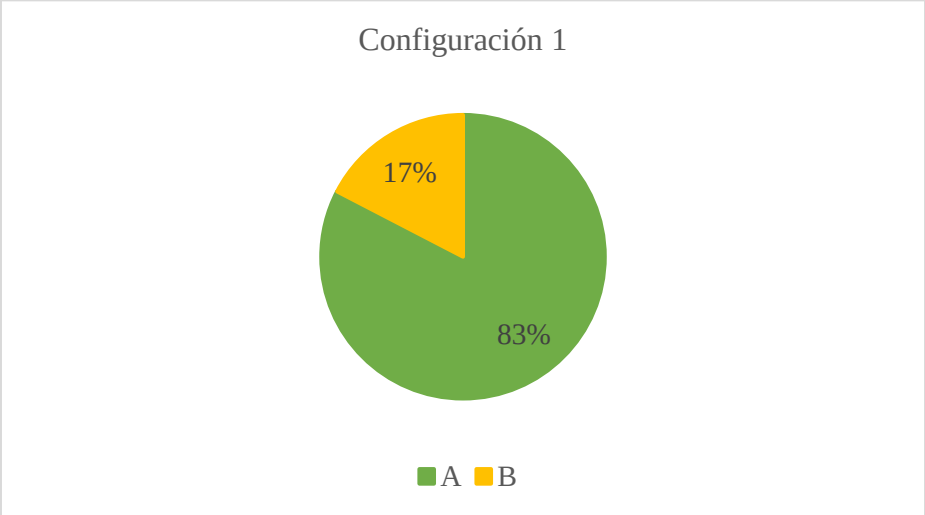


Figura 4.5. Resultados prueba ABX configuración 1.
Por los autores.

En la Figura 4.5 se observa que para la configuración 1 el 83% de los oyentes está de acuerdo en que la muestra transformada X corresponde al hablante objetivo, mientras que un 17% dice que corresponde al hablante origen.

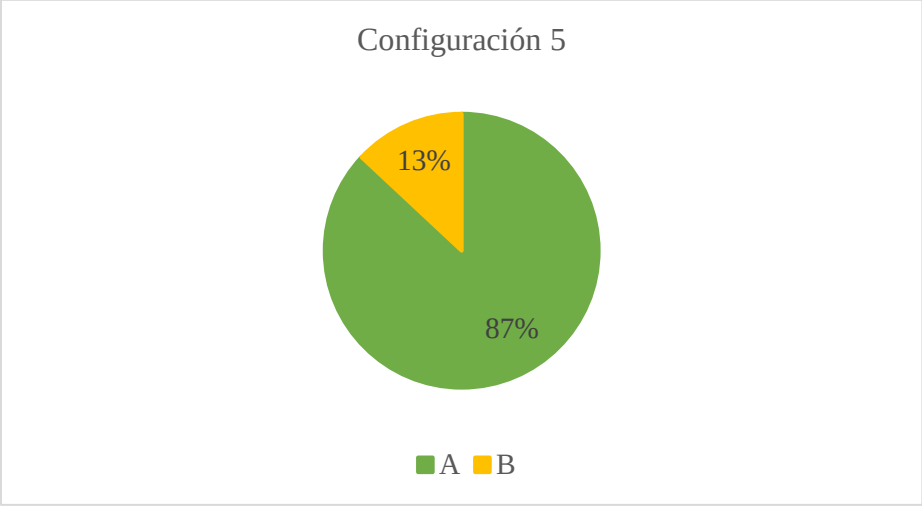


Figura 4.6. Prueba ABX configuración 5.
Por los autores.

Para la configuración 5 (ver Figura 4.6) el 87% de los oyentes están de acuerdo en que la muestra transformada corresponde al hablante objetivo y el 13% al hablante origen. Se observa que la configuración 5 tiene un incremento de los resultados, a favor del sistema, con respecto a la configuración 1.

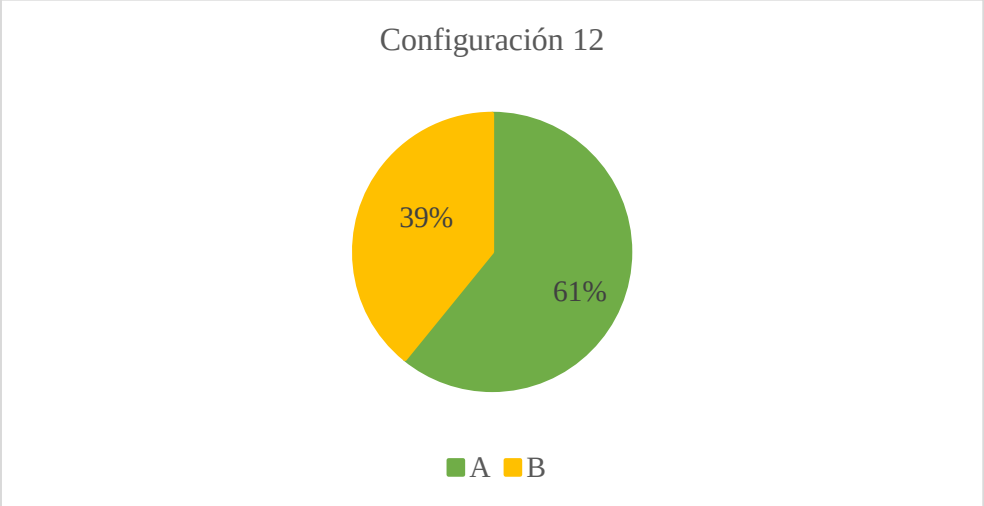


Figura 4.7. Prueba ABX configuración 12.
Por los autores.

Para la configuración 12 (ver Figura 4.7) el 61% de los oyentes coincide en que la muestra transformada X corresponde al hablante objetivo y el 39% restante al hablante origen, lo cual nos permite evidenciar que esta configuración no logra realizar una buena transformación de la identidad del hablante.

Los resultados muestran que las configuraciones 1 y 5 del sistema tienen un buen rendimiento en la transformación de la voz origen a la voz objetivo. Para estas dos configuraciones existe un mayor consenso en que la muestra transformada corresponde a la muestra objetivo, mientras que para la configuración 12 existe un porcentaje importante de oyentes que consideran que la muestra transformada corresponde a una muestra del hablante origen, lo que implica un bajo rendimiento del sistema de *Voice Morphing* con esta configuración.

MOS

Para la construcción de la prueba MOS, se establecen 4 características a evaluar, cada una con 5 posibles valoraciones: Malo (siendo ésta la peor), pobre, regular, bueno y excelente (siendo ésta la mejor).

Las características para evaluar son:

- **Problemas de comprensión**, es decir, qué tan inteligible es el mensaje que pronuncia el hablante.
- **Articulación del sonido del habla**, esto es, cómo se percibe la pronunciación del hablante.
- **Voz humana**, en otras palabras, si se percibe como una voz natural o más como una voz robotizada.
- **Impresión global**, mediante la cual se pretende conocer cómo valora el oyente la calidad de la voz transformada.

Los resultados de la prueba MOS se presentan en las siguientes gráficas.

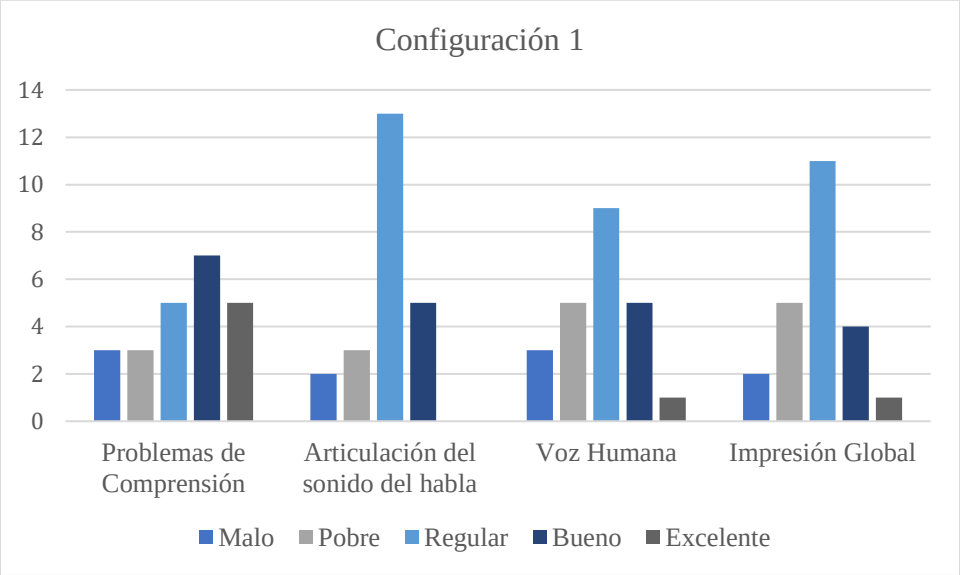


Figura 4.8. Resultados prueba MOS configuración 1.
Por los autores.

Para la configuración 1 (ver Figura 4.8) se observa en general las características evaluadas tienen un nivel regular. Los problemas de comprensión de las muestras transformadas tienen más votos en nivel excelente que las demás características evaluadas, sin embargo, la articulación del habla no tiene ningún voto en el nivel excelente. Estos resultados muestran que con esta configuración se perciben artificios que le restan naturalidad a la voz, a pesar de ello, la inteligibilidad del mensaje de la muestra no se afecta significativamente.

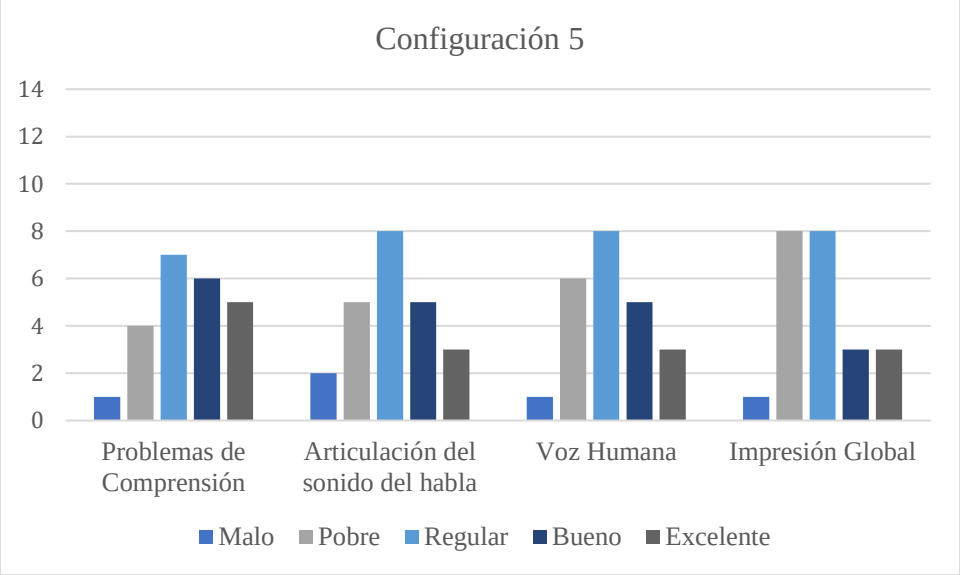


Figura 4.9. Resultados prueba MOS configuración 5.
Por los autores.

Para la configuración 5 (ver Figura 4.9) cada una de las características tiende al mismo comportamiento. De manera comparativa, se observa un mejor desempeño con respecto a la configuración 1, ya que en todas las características la valoración 'excelente' tiene más votos, mientras que la valoración 'malo' presenta una disminución de los votos. En general el nivel regular sigue siendo el más común, aunque cercano al pobre. Para la impresión global el nivel regular es igual al pobre.

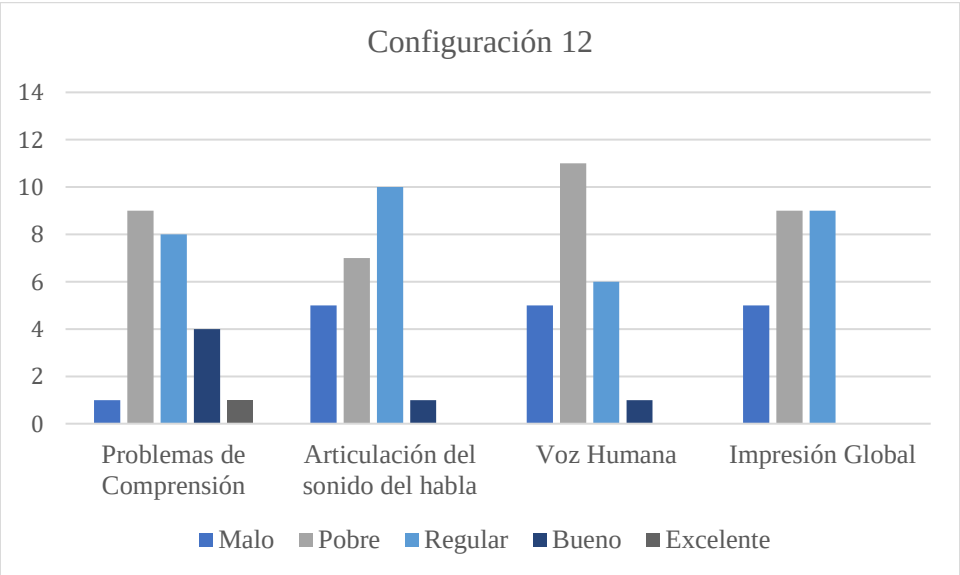


Figura 4.10. Resultados prueba MOS configuración 12.
Por los autores.

Para la configuración 12 (ver Figura 4.10) el nivel pobre tiene más votos, mientras que los niveles excelente y bueno disminuyen para cada una de las características evaluadas, siendo el primero nulo para la impresión global. Se observa además un

incremento en los votos por el nivel malo con respecto a las otras configuraciones. Estos resultados implican que para la configuración 12 la inteligibilidad del mensaje sigue siendo aceptable, sin embargo, la naturalidad de la voz y la impresión global de la muestra transformada se ven afectadas notablemente, lo que implica una calidad de audio pobre.

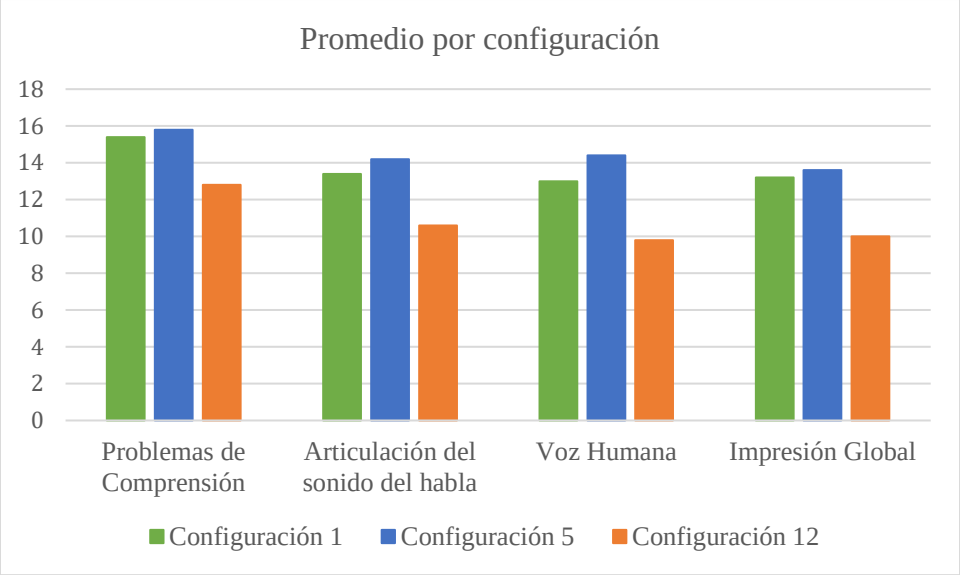


Figura 4.11. Resultados promedio prueba MOS para las 3 configuraciones.
Por los autores.

La Figura 4.11 muestra una visión general mediante los promedios de cada configuración para cada una de las características evaluadas. Se observa que la configuración 5 presenta un puntaje ligeramente superior a la configuración 1 para cada característica, mientras que la configuración 12 presenta un nivel inferior a las dos configuraciones restantes. Los problemas de comprensión tienen el puntaje más alto para cada configuración, mientras que la voz humana presenta los puntajes más bajos, junto con la impresión global, por tanto, es posible afirmar que las configuraciones empleadas en este trabajo para el proceso de *Voice Morphing,* no alteran la inteligibilidad del mensaje pronunciado por el hablante. No obstante, la naturalidad de las muestras transformadas se ve directamente afectada por la configuración y características del sistema. Adicionalmente, se puede inferir que este hecho es debido a que las señales de origen contienen algunos elementos que no corresponden al habla en sí, como por ejemplo la respiración audible, los modelos de conversión crean artefactos audibles en las muestras transformadas.

Otro aspecto que afecta la naturalidad de la voz es la conversión por tramas, que hace que las discontinuidades en la señal de voz generen ruido en las muestras transformadas.

Analizando los resultados obtenidos en la evaluación subjetiva, se puede evidenciar que no hay una concordancia total con los resultados obtenidos en la prueba

objetiva, puesto que en esta última, la configuración 1 obtuvo mejores resultados que la configuración 5, por tanto, es necesario realizar un análisis estadístico de los resultados obtenidos en las pruebas y determinar si existe una diferencia significativa de la apreciación de los oyentes entre las diferentes configuraciones, para esto se aplica una prueba ANOVA.

ANOVA

El Análisis de la Varianza (ANOVA, *Analysis of Variance*) es un método estadístico que permite analizar e interpretar observaciones de varias poblaciones. ANOVA clasifica las observaciones en contribuciones de diferentes fuentes y luego determina si existen diferencias significativas entre las fuentes de variación o no, para lo cual calcula un valor, *p*, que expresa la cantidad de esta variación [55]. Si *p* está por debajo de 0.05, se refuta la hipótesis nula a favor de la alternativa, esto quiere decir que al menos la media de uno de los grupos es significativamente diferente [56].

Existe ANOVA de una vía, cuando hay una sola variable independiente para clasificar los grupos, y de dos o más niveles. Esta última es una extensión fundamentada en el mismo razonamiento que la primera [57].

En este trabajo se utiliza ANOVA de una vía para analizar los resultados obtenidos en las pruebas subjetivas. La Tabla 4.4 presenta los resultados, valores de *p*, para cada uno de los grupos evaluados, los cuales corresponden a las 3 configuraciones evaluadas. Las variables independientes son las características evaluadas.

Tabla 4.4. Resultados prueba ANOVA entre las configuraciones 1, 5 y 12.

Parámetro	**Configuraciones**			
	12-5-1	**1-5**	**1-12**	**5-12**
ABX	0.0816	0.6895	0.106	0.045
Problemas de Comprensión	0.1253	0.8148	0.1054	0.043
Articulación del sonido del habla	0.021	0.5655	0.021	0.0134
Voz humana	0.004	0.3472	0.0172	0.0011
Impresión global	0.0136	0.778	0.0101	0.0081

Se observa que la configuración 5 y la configuración 1 presentan valores muy grandes con respecto al valor máximo de *p*, lo que muestra una variación de medias poco significativa, esto debido a que las dos configuraciones presentan resultados muy similares en la evaluación subjetiva. Es importante resaltar que tanto la

configuración 1 como 5 utilizan una matriz de covarianza completa, 24 coeficientes MFCC para representar las envolventes y difieren en el número de componentes gaussianos, 32 y 16 respectivamente.

Teniendo en cuenta estos resultados y los resultados de la prueba objetiva, en los que se observa que la configuración 1 tiene menor nivel de distorsión que la configuración 5, es posible reiterar que entre las características a variar en cada una de las configuraciones la que tiene menos efecto en la señal resultante es el número de componentes gaussianos. Sin embargo, es importante encontrar un número adecuado de componentes gaussianos a utilizar, puesto que el uso de pocos componentes puede llevar a una mala representación de las señales de voz de los hablantes, y el uso de muchos componentes, en el caso de la matriz de covarianza completa, puede llevar a un sobreajuste, además de un aumento considerable en el costo computacional. [29]. En la Figura 4.4 se observa que con la configuración 9 también se obtienen resultados deseables en la prueba objetiva, siendo la diferencia con las configuraciones 1 y 5 que el número de componentes gaussianos en este caso es 8.

Por otra parte, para la configuración 5 y la configuración 12 el valor de *p* se encuentra por debajo de 0,05 lo que indica una mayor variación de las medias entre las dos configuraciones, lo cual se cumple para cada uno de los factores analizados, reafirmando la diferencia entre estas dos configuraciones. Mientras que al comparar la configuración 1 y la configuración 12, se observa que éstas no presentan una variación significativa en la prueba ABX y en la característica de problemas de comprensión, sin embargo, para las características restantes existen variaciones representativas.

Cuando se hace una comparación entre las tres configuraciones, se observa que las principales diferencias no se encuentran en la prueba ABX y los problemas de comprensión, sino en las características con una mayor relación frente a la naturalidad de la voz, por lo cual se deduce que las configuraciones 1 y 5 permiten reducir los artefactos introducidos en el proceso de *Voice Morphing* posibilitando mejorar la naturalidad de la voz obtenida.

CAPÍTULO 5: *CONCLUSIONES Y TRABAJOS FUTUROS*

En el presente trabajo de investigación se expuso el diseño, implementación y evaluación de un sistema de *Voice Morphing* basado en transformaciones lineales. Con el propósito de evaluar el desempeño del sistema, se realizaron tanto pruebas objetivas como pruebas subjetivas. Las primeras se aplicaron a 12 configuraciones diferentes del sistema, en las cuales se varían parámetros como: el número de componentes gaussianos, número de coeficientes MFCC y tipo de matriz de covarianza. Por otro lado, las pruebas subjetivas se aplicaron a las dos mejores y a la peor configuración de acuerdo con las pruebas objetivas.

El siguiente capítulo presenta las conclusiones obtenidas a partir del desarrollo del trabajo de grado titulado: “Evaluación del rendimiento de un algoritmo de *Voice Morphing* soportado en transformaciones lineales según la naturalidad de la voz obtenida”, así como el planteamiento de algunos posibles trabajos de investigación a futuro que pueden realizarse a partir de los resultados obtenidos aquí.

5.1. CONCLUSIONES

- La selección del tipo de matriz de covarianza en la configuración del sistema de *Voice* Morphing tiene un alto impacto en la función de conversión resultante. La matriz de covarianza completa presenta mejores resultados, sin embargo, la matriz de covarianza diagonal en bloque es más eficiente desde el punto de vista computacional.

- La selección del número de componentes gaussianos no afecta significativamente el desempeño del sistema de *Voice Morphing.* Los resultados de las pruebas objetivas muestran que el mejor desempeño se obtiene con las configuraciones 1, 5 y 9, entre las cuales solo cambia el número de componentes gaussianos. De igual forma, los resultados subjetivos muestran que no existe una diferencia significativa en la percepción de la calidad de las señales de voz obtenidas con las configuraciones 1 y 5. Por lo anterior, para este caso es recomendable utilizar un número de componentes gaussianos entre 8 y 16, con el fin de no incrementar innecesariamente el costo computacional y evitar el sobreajuste.

- La información de las señales de voz que se introducen en el sistema de *Voice Morphing* se encuentra representada por medio de sus envolventes espectrales, por lo que un mayor número de coeficientes MFCC permite obtener una representación más acertada. En la literatura se encuentra que los primeros 12 coeficientes MFCC contienen la mayor parte de la información, no obstante, según los resultados obtenidos en este trabajo de grado, la calidad de la señal de voz resultante es mejor al utilizar un número de coeficientes MFCC igual a 24.

- Existen varias dificultades relacionadas con la evaluación objetiva de la calidad del sistema de *Voice Morphing*. En primer lugar, para los datos de prueba, los resultados objetivos generalmente se basan en una alineación imperfecta de las tramas a comparar. Además, no es posible modelar matemáticamente la percepción del sonido por medio del oído humano, por lo que no se tiene certeza sobre cómo se concibe la calidad general del sonido. Las pruebas subjetivas, por su parte, permiten hacer una estimación de la calidad de la voz transformada, con la cual es posible determinar la exitosa transformación de la identidad del hablante, así como también la naturalidad de la voz transformada.

- Para obtener unas muestras transformadas con una calidad adecuada y naturalidad en la voz, y además con un costo computacional razonable se requiere encontrar un equilibrio entre los parámetros de matriz de covarianza, orden de los componentes gaussianos y número de coeficientes MFCC. Es importante resaltar que la matriz de covarianza diagonal en bloque presenta un buen desempeño como base para la variación de los demás parámetros, puesto que se disminuye el costo computacional y se evita el sobreajuste que se puede presentar al usar una matriz de covarianza completa.

5.2. TRABAJOS FUTUROS

Debido a que *Voice Morphing* es un campo de estudio de interés, en el que actualmente se realizan avances en las técnicas y configuraciones de los sistemas, se pueden tomar diferentes criterios para diseñar nuevos modelos o mejorar los ya existentes. Teniendo en cuenta los resultados obtenidos en este trabajo de grado, a continuación, se enlistan propuestas para trabajos futuros.

- Aplicar el algoritmo de *Voice Morphing* cuando se tienen como hablante origen y objetivo personas de diferente género y analizar el efecto sobre la naturalidad de la voz resultante cuando se varían los parámetros del número

de coeficientes, el número de componentes gaussianos y el tipo de matriz de covarianza.

- Realizar el análisis de desempeño y el efecto sobre la naturalidad de la voz de las muestras transformadas del sistema de *Voice Morphing* cuando además de variar los parámetros del número de coeficientes, el número de componentes gaussianos, el tipo de matriz de covarianza, se varía el número de datos de entrenamiento del GMM.

- Realizar el análisis de desempeño y el efecto sobre la naturalidad de la voz de las muestras transformadas del sistema de *Voice Morphing* variando la nacionalidad y el género de las bases de datos de las personas empleadas para el entrenamiento y evaluación.

- Comparar el rendimiento y la naturalidad de la voz de las muestras transformadas del algoritmo de *Voice Morphing* implementado; basado en transformaciones lineales, con un algoritmo de *Voice Morphing* basado en redes neuronales.

REFERENCIAS

[1] Z. Zhang, “Mechanics of human voice production and control,” The journal of the acoustical society of America, vol. 140, no. 4, pp. 2614–2635, 2016.
[2] F. Miyara, “La voz humana," Laboratorio de Acústica y Electroacústica, Escuela de Ingeniería Electrónica, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina. Recuperado de: http://www.fceia.unr.edu.ar/prodivoz/fonatorio.pdf, 1999.
[3] C. Duque Sánchez and M. Morales López, et al., Caracterización de voz empleando análisis tiempo - frecuencia aplicada al reconocimiento de emociones. PhD thesis, Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación.
[4] "Ejercicio de El aparato respiratorio", Es.liveworksheets.com, 2021. [Online]. Available: https://es.liveworksheets.com/worksheets/es/Ciencias_de_la_Naturaleza/El_aparato_respiratorio/El_aparato_respiratorio_ql32820hp. [Accessed: 21- Aug- 2021].
[5] D. P. Alonso, Contribución al estudio de selección de parámetros para identificación de estrés en la voz. PhD thesis, Universidad Politécnica de Madrid, 2017.
[6] B. T. Gallardo, “La voz y nuestro cuerpo: un análisis funcional,” Revista de Investigaciones en Técnica vocal, vol. 1, pp. 40–58, 2013.
[7] B. Torres, “Anatomía funcional de la voz”. Capítulo 1 del libro: Medicina del Canto. URL: http://www.medicinadelcant.com/cast/llibre.htm#, 2007.
[8] Mintz, A. Pérez, A. Peñalosa, B. Bider, M. Chalup, and J. I. Barreras, “Fisiología de la faringe,” Rev FASO, vol. 21, pp. 27–9, 2014.
[9] Silva, M. M. (2021) *Cuantificación de señales de voz soportada en su representación Wavelet*. Unpublished manuscript.
[10] Anonymous. (2007) "File: Glottal Cycle.gif - Wikimedia Commons", Commons.wikimedia.org. [Online]. Available: https://commons.wikimedia.org/wiki/File:Glottal_Cycle.gif. [Accessed: 22- Oct- 2021].
[11] J. I. Hualde, Los sonidos del español: Spanish language edition. Cambridge University Press,2013.
[12] I. Cobeta, F. Nuñez, and S. Fernández, Patología de la voz. Marge books, 2013.
[13] J. L. Navarro Mesa, “Procesador acústico: El bloque de extracción de características,” 1994.
[14] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. I-709, doi: 10.1109/ICASSP.2004.1326084.
[15] "G.722 Wideband Audio Codec Support Across TDM and VoIP Platforms - GL Press Release", Gl.com, 2021. [Online]. Available: https://www.gl.com/press-

release/g722-wideband-audio-codec-support-across-tdm-voip-platforms-press-release.html. [Accessed: 21- Aug- 2021].
[16] S. Frühholz and P. Belin, The Oxford handbook of voice perception. Oxford University Press, 2018.
[17] I. Ahmed, A. Sadiq, M. Atif, M. Naseer and M. Adnan, "Voice morphing: An illusion or reality," *2018 International Conference on Advancements in Computational Sciences (ICACS)*, 2018, pp. 1-6, doi: 10.1109/ICACS.2018.8333282.
[18] C. Orphanidou, I. Moroz, and S. Roberts, “Wavelet-based voice morphing,” 2004.
[19] Hui Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1301-1312, July 2006, doi: 10.1109/TSA.2005.860839.
[20] C. Orphanidou, I. M. Moroz, and S. J. Roberts, “Voice morphing using the generative topo-graphic mapping,” 2003.
[21] H. Ye and S. Young, “High quality voice morphing,” in2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I–9, IEEE, 2004.
[22] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," in IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, March 1998, doi: 10.1109/89.661472.
[23] A. Kain, High-resolution voice transformation. Oregon Health & Science University, 2001.
[24] H. Ye and S. Young, “Perceptually weighted linear transformations for voice conversion,” in Eighth European Conference on Speech Communication and Technology, 2003.
[25] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad, "Voice conversion using Artificial Neural Networks," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 3893-3896, doi: 10.1109/ICASSP.2009.4960478.
[26] V. V. Das, Proceedings of the Third International Conference on Trends in Information, Telecommunication and Computing, vol. 150. Springer Science & amp; Business Media, 2012.
[27] A. Bhatt and M. Scholar, “A psola based approach for Voice Morphing,” International Journal of Digital Application and Contemporary Research (IJDACR), Feb-2015
[28] G. Martínez and G. Aguilar, “Reconocimiento de voz basado en MFCC, SBC y espectrogramas,” Universidad Politécnica Salesiana, 2013.
[29] E. Helander, “Mapping techniques for voice conversion,” 2012.
[30] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in In International Symposium on Music Information Retrieval, Citeseer, 2000.
[31] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in Proceedings of the 1998 IEEE International Conference on Acoustics,

Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 1, pp. 285–288, IEEE, 1998.
[32] "MFCC", *MathWorks*, 2021. [Online]. Available: https://la.mathworks.com/help/audio/ref/mfcc.html. [Accessed: 30- Sep- 2021].
[33] D. Yu and L. Deng, "Gaussian mixture models," in Automatic Speech Recognition, pp. 13–21, Springer, 2015.
[34] B. A. S. Hasan and J. Q. Gan, "Sequential EM for unsupervised adaptive gaussian mixture model-based classifier," in International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 96–106, Springer, 2009.
[35] T. Silva, "How to code Gaussian Mixture Models from scratch in Python", Medium, 2019. [Online]. Available: https://towardsdatascience.com/how-to-code-gaussian-mixture-models-from-scratch-in-python-9e7975df5252. [Accessed: 26- Oct- 2021].
[36] P. Zolfaghari and T. Robinson, "A formant vocoder based on mixtures of gaussians," in1997IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1575–1578, IEEE, 1997.
[37] Y. Stylianou, O. Cappe and E. Moulines, "Continuous probabilistic transform for voice conversion," in IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, March 1998, doi: 10.1109/89.661472.
[38] Hui Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1301-1312, July 2006, doi: 10.1109/TSA.2005.860839.
[39] S. J. Leon, I. Bica, and T. Hohn, Linear algebra with applications. Pearson Prentice Hall Upper Saddle River, NJ, 2006
[40] J. Z. Gamboa, "Evolución de las metodologías y modelos utilizados en el desarrollo de software," INNOVA Research Journal, vol. 3, no. 10, pp. 20–33, 2018.
[41] P. J. Sáez Martínez, V. Rodríguez Montequín, J. Villanueva Balsera, and M. Cueto Cuiñas, "Selección de modelos y metodologías agiles en proyectos software," 2014.
[42] L. Inteco, "Ingeniería del software: metodologías y ciclos de vida," Laboratorio Nacional de Calidad Del Software, vol. 83, 2009.
[43] "Festvox: CMU_ARCTIC Databases", Festvox.org, 2021. [Online]. Available: http://www.festvox.org/cmu_arctic/. [Accessed: 21- Aug- 2021].
[44] D. M. Verde and D. M. VERDE, "Análisis y reconocimiento de voz esofágica," 2005.
[45] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," arXiv preprint ar-Xiv:1003.4083, 2010.
[46] J. Benesty, M. M. Sondhi, Y. Huang, et al., Springer handbook of speech processing, vol. 1. Springer, 2008.

[47] X. Zhou, Z.-q. Yao, and B. Dai, “Improved covariance modeling for GMM in speaker identification,” in Ninth European Conference on Speech Communication and Technology, 2005.
[48] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” Digital signal processing, vol. 10, no. 1-3, pp. 19–41, 2000.
[49] J. Nurminen, V. Popa, J. Tian, Y. Tang, and I. Kiss, “A parametric approach for voice conversion,” TCSTAR WSST, pp. 225–229, 2006.
[50] E. Helander, T. Virtanen, J. Nurminen and M. Gabbouj, "Voice Conversion Using Partial Least Squares Regression," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 912-921, July 2010, doi: 10.1109/TASL.2010.2041699.
[51] M. Morise, “Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals,” in Proc. Interspeech 2017, pp. 2321–2325, 2017.
[52] M. Morise, “Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,” Speech Communication, vol. 67, pp. 1–7, 2015.
[53] M. Morise, “D4c, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol. 84, pp. 57–65, 2016.
[54] M. D. Polkosky and J. R. Lewis, “Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X,” International Journal of Speech Technology, vol. 6, no. 2, pp. 161–182, 2003.
[55] A. Salami, F. Ghassemi and M. Hasan Moradi, "A Criterion to Evaluate Feature Vectors Based on ANOVA Statistical Analysis," *2017 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, 2017, pp. 14-15, doi: 10.1109/ICBME.2017.8430266.
[56] J. Korstanje, "1-way ANOVA from scratch—dissecting the ANOVA table with a worked example", Medium, 2021. [Online]. Available: https://towardsdatascience.com/1-way-anova-from-scratch-dissecting-the-anova-table-with-a-worked-example-170f4f2e58ad. [Accessed: 30- Sep- 2021].
[57] J. Dagninoet, “Análisis de varianza,” Revista chilena de anestesia, vol. 43, no. 4, pp. 306–310, 2014.
[58] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” J. Acoust. SOC. her., vol. 57, S35(A), 1975.
[59] P. F. de Carrera, “Estudio y simulación de un codificador de voz basado en la recomendación g. 729 de la ITU-T,” 2005.
[60] F. Zheng, Z. Song, L. Li, W. Yu, F. Zheng, and W. Wu, “The distance measure for line spectrum pairs applied to speech recognition,” in Fifth International Conference on Spoken Language Processing, 1998.

APÉNDICE A

En el apéndice A se documenta con mayor detalle lo relacionado con las Frecuencias Espectrales de Línea y la función de transferencia del modelo del tracto vocal.

A.1. Frecuencia Espectrales de Línea

El concepto de Frecuencias Espectrales de Línea es introducido inicialmente por Itakura [58], y plantea una representación paramétrica alternativa de los LPC.

La idea fundamental del análisis LPC es que la muestra de la señal actual se pueda aproximar mediante una combinación lineal de las muestras anteriores [59]:

$$s(n) \approx \sum_{k=1}^{p} a_k s(n-k)$$

En la compresión y el reconocimiento de voz fundamentado en la codificación se conoce que los coeficientes LPC $\{a_1, a_2, \dots, a_p\}$ son inadecuados debido a posibles problemas de inestabilidad del filtro. Con el objetivo de mitigar estos problemas se planteó un conjunto diferente de parámetros que representan la misma información espectral, como por ejemplo los coeficientes de reflexión y relaciones de área logarítmica, etc., para la cuantificación. LSF hace parte de estos tipos de representación de información espectral; estos parámetros tienen propiedad de preservación de estabilidad del filtro y se pueden emplear para codificar información espectral LPC de manera más eficiente que cualquier otro parámetro [60].

La frecuencia central de un formante es caracterizada por dos o tres parámetros LSF y su ancho de banda depende de la proximidad de los respectivos parámetros Figura 2. 3 [59].

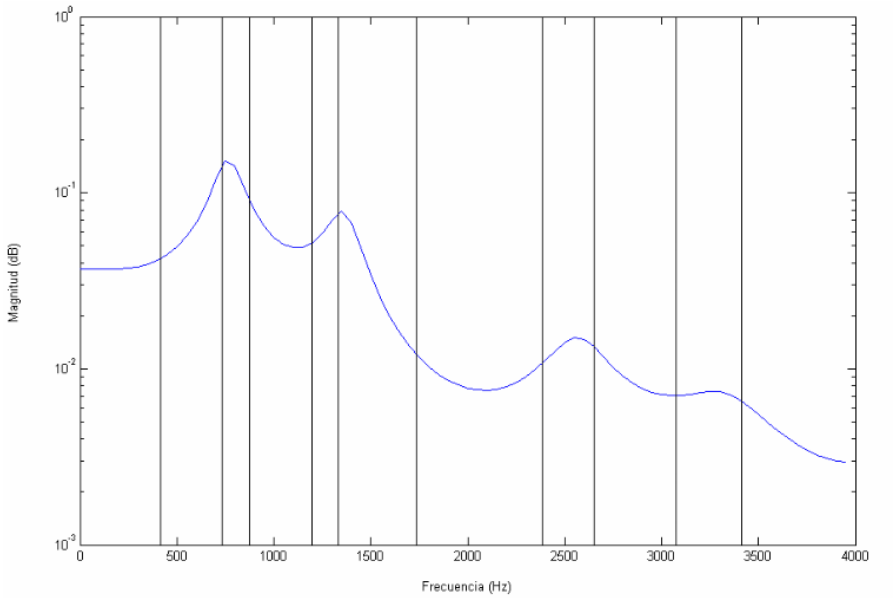


Figura A. 1. Espectro de frecuencia de un tramo de voz con la posición de los coeficientes LSF. Adaptado de [59].

Una característica importante para resaltar es que durante los fonemas sonoros los parámetros LSF no varían considerablemente, mientras que durante los fonemas sordos pueden variar rápidamente. Tanto el espectro de la señal de voz como los parámetros LSF tienen segmentos redundantes al igual que variaciones representativas, esto se puede observar sobre todo en los parámetros LSF más bajos, los cuales durante cientos de milisegundos se sostienen en el mismo estado y luego cambian repentinamente.

Otra singularidad relevante de los parámetros LSF es su sensibilidad espectral localizada, es decir, que pequeñas modificaciones en los parámetros LSF producen un cambio en la respuesta en magnitud solamente en las frecuencias cercanas a la frecuencia del parámetro modificado, lo que hace a las LSF muy tolerantes a los errores, en comparación con otras representaciones [59].

APÉNDICE B

En el apéndice B se presenta las fuentes del código del proyecto Sprocket, así como el código de MATLAB utilizado para la evaluación objetiva y las bases de datos utilizadas para la implementación del sistema de *Voice Morphing*.

B.1. Proyecto Sprocket

El código correspondiente al proyecto base se encuentra en la plataforma de desarrollo colaborativo GitHub:
https://github.com/k2kobayashi/sprocket

B.2. Evaluación Objetiva

El código para la evaluación objetiva se muestra a continuación:

```
clc
clear
files = dir('*.h5');
k=0;
arregloPromMcd = [];
arregloPromRmse = [];
for i = 1:2:70
      k = i+1;
      h2 = files(i).name;
      h1 = files(k).name;
      data = hdf5info(h1);
      coeffsH1 = hdf5read(data.GroupHierarchy.Datasets(3));
      data2 = hdf5info(h2);
      coeffsH2 = hdf5read(data2.GroupHierarchy.Datasets(3));
    %Alineamiento DTW de tramas MFCC Origen y Destino
    [ds,ix,iy] = dtw(coeffsH1,coeffsH2);
    coeffsH1Alineados = coeffsH1(:,ix);
    coeffsH2Alineados = coeffsH2(:,iy);
    [n,m]=size(coeffsH1Alineados);
    tamanoVector = n*m;
    %Cálculo MCD
    diferenciaMcd = coeffsH1Alineados - coeffsH2Alineados;
    cuadradoMcd = diferenciaMcd.^2;
    sumatoriaMcd = sum(cuadradoMcd);
    doble = sumatoriaMcd.*2;
    mcdTramas = (10/log(10))*sqrt(doble);
    promedioMcd = mean(mcdTramas);
    arregloPromMcd(i) = promedioMcd;
    %Cálculo RMSE
    diferenciaRmse = coeffsH1Alineados - coeffsH2Alineados;
```

```
      cuadradoRmse = diferenciaRmse.^2;
      sumatoriaRmse = sum(cuadradoRmse);
      fracción = sumatoriaRmse.*(1/n);
      raíz = sqrt(fraccion);
      promedioRmse = mean(raiz);
      arregloPromRmse(i) = promedioRmse;
end
mcdTotal = mean(arregloPromMcd);
RmseTotal = mean(arregloPromRmse);
```

B.3. Bases de Datos

Las bases de datos utilizadas para la implementación del sistema de *Voice Morphing* son las bases de datos para síntesis de voz CMU_ARCTIC. Las bases de datos están conformadas por alrededor de 1150 oraciones seleccionadas de textos sin derechos de autor del Proyecto Gutenberg. Estas incluyen hablantes masculinos y femeninos de inglés estadounidense y otros acentos como el escocés y el hindú.

Las bases de datos están disponibles en:
http://www.festvox.org/cmu_arctic/

El informe detallado de la estructura y el contenido de las bases de datos se encuentra en el documento:
http://www.festvox.org/cmu_arctic/cmu_arctic_report.pdf

APÉNDICE C

En el apéndice B se presenta el formulario utilizado para las pruebas subjetivas del sistema de *Voice Morphing*.

Se diseñaron 3 formularios con el objetivo de utilizar distintas muestras transformadas para cada una de las configuraciones evaluadas. A continuación, se presenta el diseño del Formulario 1. Los formularios 2 y 3 siguen el mismo diseño y presentación de las preguntas para cada una de las pruebas.

C.1. Prueba ABX

Para la prueba ABX se presentan 3 preguntas para la configuración 1, 5 y 12 respectivamente.

En cada pregunta se presentan 3 audios. Los dos primeros audios corresponden a muestras del hablante Origen y Objetivo, definidos aleatoriamente como Audio A y Audio B. La muestra transformada se denomina Audio X (ver Figura C. 1). Se le pide al oyente escuchar atentamente cada uno de los audios y seleccionar cuál de los dos Audios (A o B) se acerca más al Audio X.

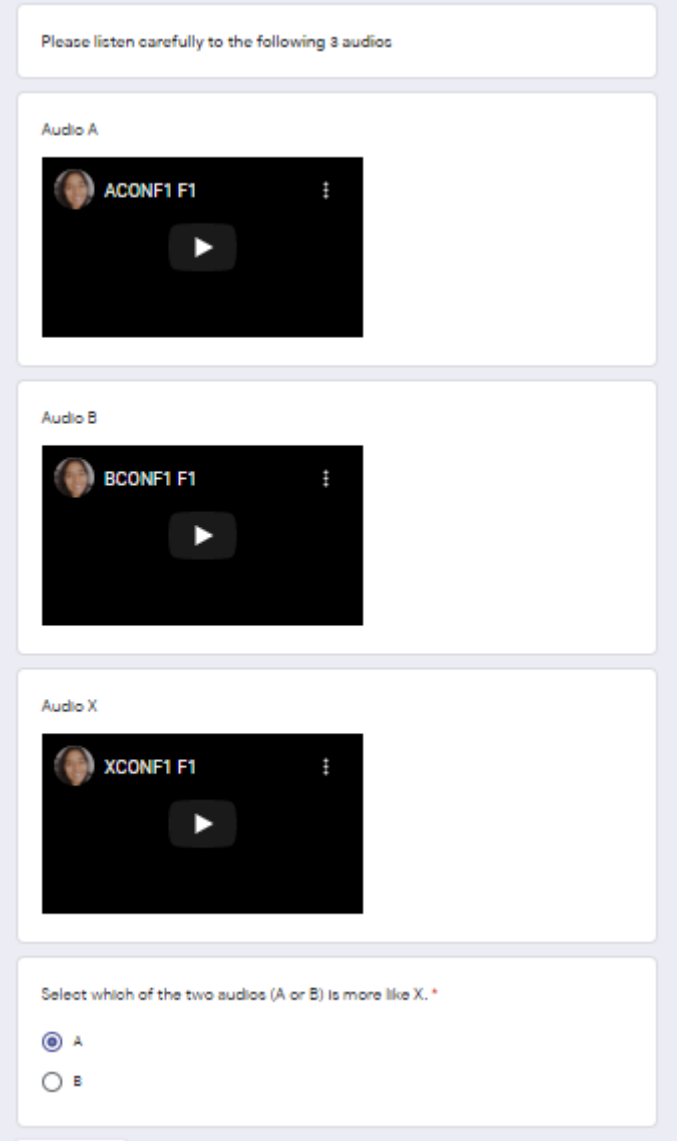


Figura C. 1. Prueba ABX para la Configuración 1.
Por los autores.

C.2. Prueba MOS

Para la prueba MOS se presentan 3 preguntas para las configuraciones 1,5 y 12 respectivamente. Las 3 preguntas se presentan en orden aleatorio para cada uno de los formularios.

En cada pregunta se le pide al oyente que escuche una muestra convertida y que califique la calidad entre (Malo, Pobre, Regular, Bueno y Excelente) de los Problemas de Comprensión, la Articulación de los sonidos del habla, la Voz Humana e la Impresión Global que presenta la muestra; como se muestra en la Figura C. 2.

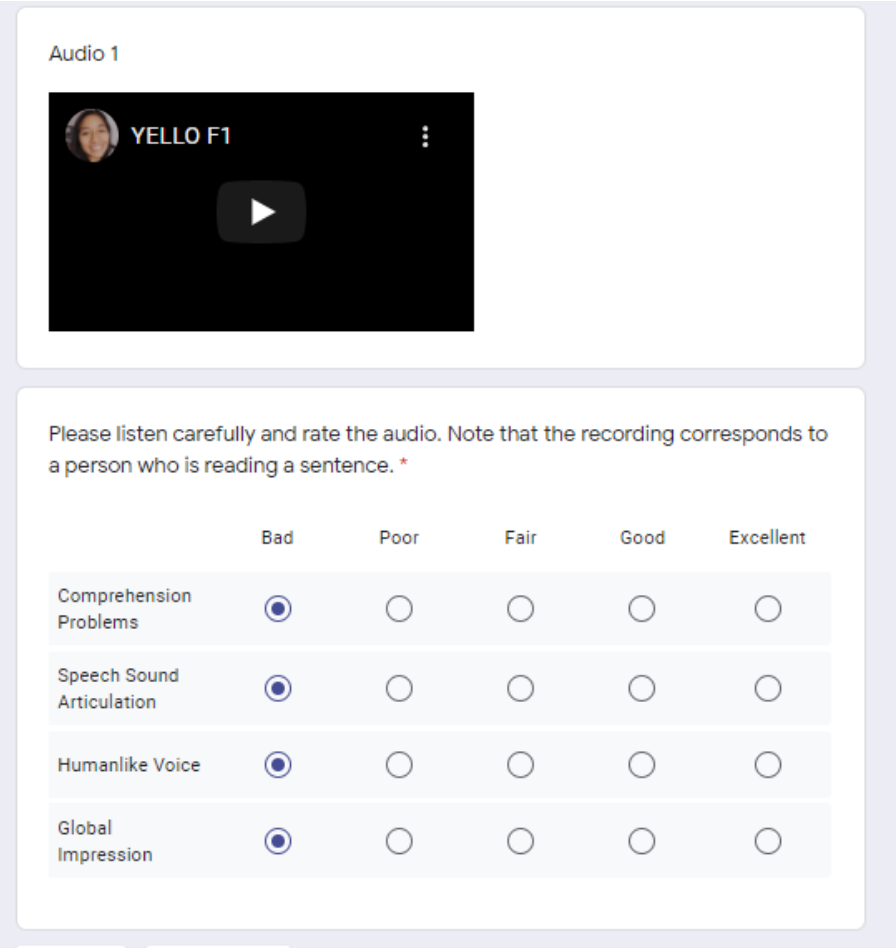
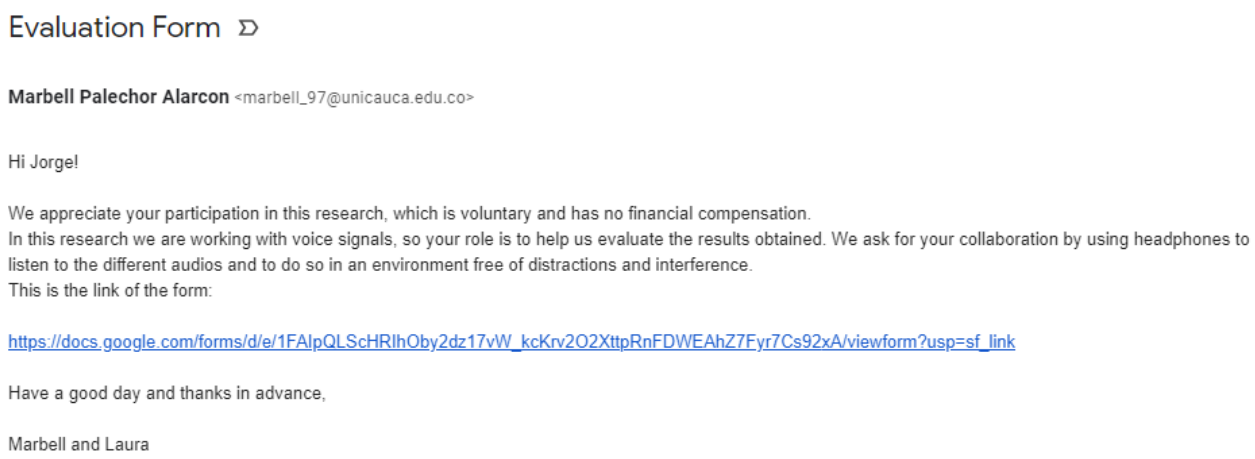


Figura C. 2. Prueba MOS para la Configuración 1.
Por los autores.

C.3. Consentimiento

En la Figura C. 3. se presenta uno de los correos electrónicos enviados a los oyentes entrevistados. En el correo se aclara que la participación es voluntaria y que no existe compensación económica. Además, se pide a los oyentes utilizar audífonos y estar en un ambiente libre de distracciones e interferencias al momento de responder el formulario. Finalmente se adjunta el enlace del formulario desarrollado en *Google Forms*.



Evaluation Form

Marbell Palechor Alarcon <marbell_97@unicauca.edu.co>

Hi Jorge!

We appreciate your participation in this research, which is voluntary and has no financial compensation.
In this research we are working with voice signals, so your role is to help us evaluate the results obtained. We ask for your collaboration by using headphones to listen to the different audios and to do so in an environment free of distractions and interference.
This is the link of the form:

https://docs.google.com/forms/d/e/1FAIpQLScHRIhOby2dz17vW_kcKrv2O2XttpRnFDWEAhZ7Fyr7Cs92xA/viewform?usp=sf_link

Have a good day and thanks in advance,

Marbell and Laura

Figura C. 3. Ejemplo del Correo Electrónico.
Por los autores.