
Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

# Implementación y evaluación de un sistema de síntesis de habla con acento extranjero variable

Tesis de Licenciatura en Ciencias de la Computación

Franco Negri

Director: Agustín Gravano

Buenos Aires, 2018

# IMPLEMENTACÍÓN Y EVALUACIÓN DE UN SISTEMA DE SINTESIS DE HABLA

El siguiente trabajo tiene como objetivo fundamental generar un sistema de síntesis de habla en castellano con acento inglés. Si bien actualmente los métodos mas avanzados para generar habla utilizan técnicas de aprendizaje automático basados redes neuronales, en este trabajo utilizaremos técnicas basadas en HMM+GMM por su practicidad. Utilizando esta técnica sintetizaremos modelos en castellano y en inglés y luego, interpolando sus características acústicas, intentaremos producir un modelo capaz de sintetizar habla que resulte inteligible pero que contenga atribuibles a un extranjero angloparlante hablando castellano. Para este trabajo la interpolación de características acústicas ocurrirá a nivel de fono, por lo que, para realizar la mezcla de modelos será necesario compatibilizar los fonos de ambos idiomas. Por este motivo, parte del trabajo consistirá en desarrollar un mapeo entre los repertorios fonéticos del inglés y el castellano que resulte natural. Por último intentaremos evaluar el nivel de efectividad de nuestro método de manera experimental. Para ello desarrollaremos una metodología de evaluación que nos permita medir tanto la ininteligibilidad de un audio, como el origen atribuido por los participantes.

**Keywords:** Síntesis de habla, HMM, HTS, GMM, acento extranjero, aprendizaje automático.

# AGRADECIMIENTOS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce sapien ipsum, aliquet eget convallis at, adipiscing non odio. Donec porttitor tincidunt cursus. In tellus dui, varius sed scelerisque faucibus, sagittis non magna. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Mauris et luctus justo. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Mauris sit amet purus massa, sed sodales justo. Mauris id mi sed orci porttitor dictum. Donec vitae mi non leo consectetur tempus vel et sapien. Curabitur enim quam, sollicitudin id iaculis id, congue euismod diam. Sed in eros nec urna lacinia porttitor ut vitae nulla. Ut mattis, erat et laoreet feugiat, lacus urna hendrerit nisi, at tincidunt dui justo at felis. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Ut iaculis euismod magna et consequat. Mauris eu augue in ipsum elementum dictum. Sed accumsan, velit vel vehicula dignissim, nibh tellus consequat metus, vel fringilla neque dolor in dolor. Aliquam ac justo ut lectus iaculis pharetra vitae sed turpis. Aliquam pulvinar lorem vel ipsum auctor et hendrerit nisl molestie. Donec id felis nec ante placerat vehicula. Sed lacus risus, aliquet vel facilisis eu, placerat vitae augue.

*A Agustin Gravano: Por tolerar mis faltas de ortografía*

# Índice general

# 1. INTRODUCCIÓN

Un sistema de *Text To Speech* (TTS) es aquel que genera habla artificial a partir de un texto de entrada. En la actualidad estos sistemas se encuentran incluidos en diversas aplicaciones domesticas: navegación por GPS, asistentes personales inteligentes (como es el caso de SIRI), ayuda para personas no videntes, y traducción automática, entre otros.

En las últimas décadas se han visto grandes progresos en este campo. Algunos sistemas son capaces de modelar con cierto grado de efectividad cuestiones tales como el acento, el tono y la entonación de un hablante (es decir, su *prosodia*). Por otro lado también se ha visto un sustancial progreso en modelar emociones y cuestiones que pueden darnos a entender que una oración es una pregunta, una orden, etc [3] [4].

En esta tesis se estudia una manera de sintetizar habla en español con acento extranjero. Existen muchos motivos por los que podría querer construirse un sistema con estas características. Por ejemplo en investigaciones de carácter lingüístico, podría querer utilizarse para vislumbrar limites en que un acento deja de parecernos local para pasar a sonar extranjero. Por otro lado en investigaciones de carácter psicológico, podría querer utilizarse para medir la confianza que deposita un oyente sobre hablantes de distinta nacionalidad. Por ejemplo, al pedir indicaciones de cómo llegar a un lugar es lógico pensar que uno no depositaría el mismo nivel de confianza si la persona que responde suena oriundo de la localidad que a alguien que suena extranjero. Por último un sistema así podría quererse construir por temas puramente técnicos ya que permitiría generar una voz en castellano, por ejemplo, combinando una voz previamente construida en algún otro idioma y un pequeño corpus de datos en castellano.

Actualmente se considera que el estado del arte para la síntesis de voz es el entrenamiento con redes neuronales profundas [1] [2]. Aún así para este trabajo decidimos utilizar *Modelos Ocultos de Markov* mas modelos de *Mezcla de Gaussianas* (HMM+GMM) que, a partir de un corpus de datos de entrenamiento, extrae información acústica y genera un modelo probabilístico que permite sintetizar habla. Para esto utilizaremos HTS [6], un framework de entrenamiento y síntesis de voz basado en HMM+GMM.

Consideramos que este método, si bien no nos permitirá obtener la mejor voz posible, es una tecnología madura y estable, con software libre disponible y de sencillo empleo, que nos permite realizar nuestros experimentos con relativa facilidad.

Como principal fuente de información utilizaremos la disertación doctoral *"Simultaneous Modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems"* del Profesor Tadashi Kitamura, Nagoya Institute of Technology [11], donde se describen de manera detallada las decisiones de diseño utilizadas en HTS, así como el modelado de parámetros para la construcción de HMMs y el modelado de cada *fono* utilizando Mel-Cepstral.

A modo de mapa conceptual, finalizamos esta introducción con un mapa esquemático de los temas centrales que se abordan en este trabajo. Primero, en la Sección 2.1 se presenta las técnicas utilizadas para el etiquetado fonético de los distintos corpus sobre los que trabajaremos. En la Sección 2.2 se presenta un mapeo entre los fonos del castellano y los del inglés, necesarios en el próximo paso. En la Sección 2.3 y 2.4 presentamos las herramientas provistas por HTS para mezclar modelos de diferentes hablantes y sintetizar habla con distintos grados de mezcla fonética y prosódica de inglés-castellano. En la Sección 2.5 se

detallan aspectos técnicos del trabajo, así cómo especificaciones de cómo modela el audio HTS, estructuras utilizadas durante el entrenamiento y otras herramientas utilizadas. Por último, en la Sección 3, intentamos validar que el sistema construido realmente cumple con las características deseadas, exponiéndose los aspectos metodológicos de la evaluación y los resultados obtenidos.

# 2. MATERIALES Y MÉTODOS

## 2.1. Preparación de los datos

El objetivo de esta sección es dar una introducción a los corpus utilizados, junto con la teoría referente a las metodologías utilizadas para su procesamiento. Los detalles implementativos serán descriptos más adelante en la Sección 2.5.

### 2.1.1. Primer corpus en castellano

Al inicio de la investigación, comenzamos con el corpus *SECYT-mujer*, construido por el Laboratorio de Investigaciones Sensoriales (INIGEM, UBA-CONICET) [7]. El mismo está compuesto por 741 oraciones cortas declarativas habladas por una locutora profesional en castellano rioplatense, equivalentes a 48 minutos de habla. A continuación, tres ejemplos de las oraciones articuladas por la locutora:

*La voluntad del juez fue impuesta en tribunales.*
*Los vinos uruguayos han mejorado en el último lustro.*
*Al atardecer se puso su disfraz juglaresco.*

Como parte inicial del trabajo es indispensable construir el etiquetado fonético para el corpus. Estos etiquetados consistirán en una lista de fonos donde se indica cuándo comienza y termina cada uno en el audio. La calidad de las oraciones que logremos sintetizar a posteriori dependerá fuertemente del etiquetado, por lo que es necesario prestar especial cuidado en que las transcripciones estén lo mejor alineadamente posible con los audios. Las transcripciones serán necesarias para entrenar los modelos de HMM+GMM, extrayendo tanto la información acústica para cada fono (cosas tales como la frecuencia principal, la duración, etc), como así también información contextual (por ejemplo: cómo suena un fono cuando está seguido de algún otro, al principio de una oración, si se encuentra en un diptongo, etc). Por ende, una mala transcripción se traducirá indefectiblemente en un mal modelo y malas oraciones sintetizadas. Llevamos a cabo varias pruebas de concepto utilizando HTS en este corpus, experimentando con diversos métodos para obtener el etiquetado fonético.

La primera estrategia consistió en utilizar alineamiento automático con *EHMM alignment* [14] empleando Festival y Festvox. Esta técnica tienen como principal ventaja su sencillez. Para generar las transcripciones solo es necesario contar con el corpus que se quiere anotar y sus *transcripciones grafémicas* (Transcripciones gráficas, escritas).

Los resultados preliminares con este método fueron bastante negativos: los audios generados resultaban poco inteligibles notándose claros defectos acústicos, el más notable siendo el fono [r] (*perro*) que se asemejaba más a [ɾ] (*pero*).

Utilizando la herramienta de visualización y manipulación de audio Praat [5] para visualizar el alineamiento entre las transcripciones fonéticas y los audios, descubrimos que la alineación de algunas oraciones estaba desfasada algunas milésimas de segundo respecto de los audios. Dado que para el alineamiento automático es necesario alinear los audios del corpus con audios sintetizados con Festival, sospechamos que pudo haber problemas con la calidad del corpus o que los audio sintetizados por Festival eran demasiado disimiles comparados con los audios reales.

Dado que para el corpus contamos con las transcripciones fonéticas anotadas de manera manual, procedimos a implementar un híbrido, que consiste en tomar los tiempos de las anotaciones fonéticas manuales y la información contextual y el repertorio fonético generado a partir del proceso de EHMM y combinarlos en una nueva transcripción fonética. De esta manera buscamos mejorar la alineación, utilizando en teoría una transcripción más precisa, pero manteniendo el repertorio fonético y la misma meta-información brindada por el alineamiento automático. Mantener el mismo repertorio fonético y meta-información de las transcripciones será clave para la síntesis de audios. Esto nos permitirá utilizar Festival para generar las "partituras" que queramos sintetizar y que estas tengan los mismos símbolos fonéticos (Para más información ver Sección 2.2).

El modelo generado con estas transcripciones mixtas resultó superior a los generadas solo con alineamiento automático. Aún así los audios sintetizados todavía no alcanzaron una calidad aceptable, realizando pruebas notamos que el sonido resultaba metálico y las frases poco inteligibles. Además se pudo percibir de manera informal otros detalles tales como que la voz original tenía un pitch mayor que la producida por los modelos, alrededor de un 10 %.

Para intentar mejorar la calidad de los audios en este punto sumamos otro corpus de datos en castellano.

### 2.1.2. Segundo corpus en castellano

En este punto de la investigación obtenemos un segundo corpus de datos, también contribuido por el Laboratorio de Investigaciones Sensoriales, *loc1_pal* [8] con 1593 oraciones cortas con una mezcla entre frases declarativas e interrogativas del 80 % y 20 % aproximadamente, pronunciadas por una locutora profesional con acento rioplatense con aproximadamente 2 horas y 26 minutos de habla. Con este nuevo conjunto de datos esperamos conseguir mejores resultados.

Presentamos tres ejemplos de las oraciones articuladas por la locutora:

*Alvarez se había animado a contarle un chiste*
*Alzó la voz para ahuyentar a los perros.*
*Ayer el general cumplió ochenta años.*

También vale la pena aclarar que algunas de las oraciones del primer y el segundo corpus están duplicadas. Por ejemplo la primera oración de ejemplo también estaba presente en *secyt-mujer*.

Para este corpus no contábamos con transcripciones fonéticas manuales por lo que nos vimos forzados a utilizar EHMM nuevamente. Aún así, los resultados fueron superiores a los conseguidos con *secyt-mujer*. Al contrario que en *secyt-mujer*, al visualizar este corpus con Praat, no pudieron apreciarse mayores desfasajes entre las transcripciones fonéticas y los audios.

Además los audios sintetizados resultaban inteligibles y con un marcado acento rioplatense. Tras algunas pruebas de concepto donde se experimentó con varios parámetros modificables dentro de HTK, logramos obtener resultados que superaban de manera significativa aquellos obtenidos previamente con *secyt-mujer*. Por consiguiente consideramos que los audios generados habían alcanzado una buena calidad que resultara ininteligible y aceptable para el objetivo de la investigación, por lo que decidimos utilizar uno de estos modelos para el resto del trabajo.

Especulamos que la disparidad en la calidad de los resultados es causada principalmente por la cantidad de audios y horas de habla de cada corpus [11]. Consideramos que esto

juega un papel predominante en la calidad de los sistemas TTS generados, aún cuando se utiliza un método de etiquetado puramente automático y propenso a errores sistemáticos en el alineamiento.

### 2.1.3. Corpus en inglés

Por otro lado entrenamos una voz en inglés utilizando el corpus *CMU-ARCTIC-SLT* [9] con 1132 oraciones en inglés y 56 minutos de habla articuladas por una mujer estadounidense, disponible en la página de HTS [6]. Ya que este corpus venía a modo de demo con HTS, asumimos que los parámetros de entrenamiento y las transcripciones fonéticas habían sido seleccionadas de manera apropiada, por lo que no intentamos mejorar la calidad de las transcripciones fonéticas más allá de lo que los scripts ofrecían.

Una vez más realizamos pruebas perceptuales informales para tener una idea cualitativa de los modelos generados. Los audios en inglés sintetizados con este modelo resultan inteligibles y con buena pronunciación. Sin embargo, en oraciones extensas, por ejemplo en audios de 17 segundos, pueden empezar a notarse errores más grabes en la prosodia. La misma se torna monótona y con ella, la inteligibilidad de los audios disminuye.

Presentamos tres ejemplos de las oraciones articuladas por la locutora:

*Author of the danger trail, Philip Steels, etc.*

*Not at this particular case, Tom, apologized Whittemore.*

*For the twentieth time that evening the two men shook hands.*

En la Sección 2.5.3 detallaremos más detenidamente los aspectos técnicos de los corpus utilizados.

## 2.2. Repertorio fonético y mapeo de fonos

Como ya adelantamos brevemente en la sección anterior, utilizamos Festvox y Festival para la generación de "partituras" necesarias en la síntesis de audios. A diferencia de las transcripciones fonéticas, estas transcripciones en realidad no se corresponden con ningún audio y son utilizadas por HTS para sintetizar habla. De alguna manera, actúan como una partitura para la generación acústica, indicándole a HTS qué fono debe sintetizar en cada momento. Es importante notar también, que los símbolos en las partituras y en las transcripciones utilizadas en el entrenamiento deben coincidir para que HTS pueda interpretarlas correctamente.

Para la posterior combinación de modelos, se presenta aquí un desafío, ya que los repertorios fonéticos de estas partituras no tienen una correspondencia directa entre inglés y castellano. Por ejemplo con el repertorio fonético de Festvox en castellano, existen tres fonos distintos para la [i], decisión que proviene de la necesidad de poder diferenciar la [i] acentuada de la no acentuada y de aquella presente en los diptongos [ia], [ie], [io], [iu]. La Tabla 2.1 muestra los repertorios utilizados por Festvox para la generación de transcripciones fonéticas de castellano (con 31 fonos) e inglés (con 40 fonos).

| Castellano | |
|---|---|
| a | m |
| a1 | n |
| b | ny |
| ch | o |
| d | o1 |
| e | p |
| e1 | r |
| f | rr |
| g | s |
| i | sp |
| i0 | t |
| i1 | th |
| k | u |
| l | u0 |
| ll | u1 |
| | x |

| Inglés | |
|---|---|
| aa | jh |
| ae | k |
| ah | l |
| ao | m |
| aw | n |
| ax | ng |
| ay | ow |
| b | oy |
| ch | p |
| d | r |
| s | sh |
| dh | t |
| eh | th |
| er | uh |
| ey | uw |
| f | v |
| g | w |
| hh | y |
| ih | z |
| iy | zh |

*Tab. 2.1:* Repertorios fonéticos utilizados por Festvox para inglés y castellano

Como solución general, ideamos el siguiente método. Tomamos los fonos generados por Festival para el corpus en inglés y los traducimos a fonos del castellano antes del entrenamiento. Luego, entrenamos el modelo en inglés, utilizando los símbolos en castellano. De esta manera, esperamos construir un modelo en inglés capaz de comprender partituras que utilizan los símbolos en castellano.

El objetivo de esta sección será entonces generar una traducción de los fonos del inglés al castellano que nos permita junto con una partitura en castellano y un corpus en inglés, sintetizar habla en castellano.

Para ello, confeccionamos una tabla donde cada fono del repertorio del castellano estuviera mapeado a uno del repertorio del inglés. Por ejemplo el fono [ae] (*alice*) del inglés lo traducimos al fono [a] (*amigo*) del castellano. De la misma manera, el fono [ao] del inglés (*for*) pasa a ser el fono [o] (*gato*) del castellano. En la Tabla 2.2 se presenta una lista exhaustiva de los reemplazos fonéticos utilizados.

*Tab. 2.2:* Mapeo Fonético

| Inglés | Castellano | Inglés | Castellano |
|---|---|---|---|
| ae | a | p | p |
| aa | a1 | r | r/rr |
| b | b | s | s |
| ch | ch | t | t |
| d | d | uw | u |
| dh | d | w | u0 |
| eh | e | uh | u1 |
| el | e1 | g | - |
| f | f | dx | - |
| hh | g [1] | em | - |
| iy | i | en | - |
| ih | i1 | er | - |
| k | k | ei | - |
| l | l | hv | - |
| jh | ll | ng | - |
| m | m | th | - |
| n | n | v | - |
| nx | n | y | - |
| n + i | ny | sh | - |
| ao | o | zh | - |
| ou | o1 | z | - |

Por otro lado, para varios fonos tuvimos que hacer reglas especiales ya que no contábamos con ningún fono del inglés lo suficientemente similar. Así, para el fono ny (ñ o ɲ, en ipa) colapsamos las apariciones del fono [n] seguido de [i]. Si bien esta solución puede parecer algo forzada, ya que estamos generando de manera casera un fono a partir de otros dos, consideramos que esto se aproxima en cierta medida a la manera real en la que un hablante no nativo aprende un idioma con una carga fonética diferente a la suya. Citando un extracto del trabajo *Transcription of Spanish and Spanish-Influenced English, Brian Goldstein, Temple University* [15]:

***Consonants***

As indicated in Table 5, there are many ways in which the features of Spanish influence the production of consonants in English. These influences cut across all sound classes, although the majority of influences will be in the fricative sound class. Several factors influence the extent to which one phonological system influences another. First, the influence may be due to the absence of phonemes or allophones in a language (Iglesias & Goldstein, 1998). For example, [$p^h$], [$t^h$], and [$k^h$] do not occur in Spanish, and [ʃ], [v], and [dʒ] do not occur in most dialects of Spanish. In attempting to produce sounds in English that do not exist in Spanish, a native Spanish speaker might substitute a close relation. Thus, /ʃ/ might be produced as [ʧ]; /ʃo/ *show* → [ʧo]. Second, there are differences in the phonotactic constraints of the two languages. In Spanish, word-initial clusters cannot begin with /s/. Thus, Spanish speakers attempting to produce English clusters of that type might exhibit either *cluster reduction* (e.g., /stɑːz/ *stars* → [tɑːz]) or *epenthesis* (or *prothesis*) (e.g., /stɑːz/ *stars* → [estɑːz]) (Perez, 1994). Third, there are differences in the distribution of sounds. In Spanish, for example, the only word-final consonants are /s/, /n/, /ɾ/, /l/, and /d/.

[1] El mapeo de [hh] a [g] en castellano resultó ser incorrecto. El fono [g] existe tanto en castellano (*gato*) como en inglés (*glad*). Notamos este error recién al hacer la evaluación perceptual, como se describe en la Sección 3.6.

Es decir, visto desde nuestra perspectiva, una persona que aprende una nueva lengua, realiza una aproximación entre los fonos conocidos y los fonos 'objetivo' de la nueva lengua. Esta perspectiva nos alienta a realizar mapeos que no resultan del todo exactos, mapeando por ejemplo el fono [w] (*twentieth*) del inglés al fono [u0] (*cuatrimestre*) del castellano o el fono [uw] (*two*) por el fono [u] (*cumplió*).

Un mapeo que podría resultar controversial es el del fono [jh] (*danger*) del inglés por el [ll] (*billete*) del castellano, porque existen otras posibilidades: podría mapearse a [sh] (ash), por ejemplo. Sin embargo, consideramos que el mapeo elegido es el que más se acerca a la pronunciación de las locutoras de las grabaciones.

De manera similar, el inglés carece del fono vibrante múltiple alveolar sordo [r] (*perro*) y dado que el fono [ɾ](*pero*) ya está siendo utilizado, no podemos mapearlo a este. Es importante reiterar aquí que no podemos dejar sin cubrir ningún fono del castellano ya que entonces el modelo generado no sería completamente compatible con la partitura en castellano.

Para hacer hincapié en el problema, supongamos que el fono [r] queda sin cubrir por ningún fono del inglés. Entrenamos el modelo con el corpus de inglés, reemplazando los símbolos del inglés por los del castellano. Ahora le damos para sintetizar una partitura con símbolos en castellano donde se encuentra presente el fono [r]. El modelo no tendrá manera de inferir el sonido correcto para este símbolo, resultando en una síntesis errónea o de baja calidad.

Una solución que encontramos para este problema es mapear cada ocurrencia de [ɾ] en inglés con equiprobabilidad a [ɾ] o a [r] en castellano. De esta manera ambos símbolos del castellano quedan cubiertos por el símbolo más similar del inglés.

Aquellos fonos del inglés que consideramos suficientemente disímiles del castellano, como es el caso de la [sh] y [z] los mapeamos a caracteres que no interfirieran para el entrenamiento ya que no los utilizaremos para la síntesis de oraciones en castellano.

Con este mapeo, podemos utilizar el corpus en inglés para sintetizar oraciones con partituras en castellano. Sin embargo, como es de esperar, los audios sintetizados resultan incomprensibles, por cuestiones como que las combinaciones de fonos del inglés y el castellano son muy distintas, así como las reglas prosódicas y las acentuaciones de las palabras, entre otras marcadas diferencias.

## 2.3. Interpolación entre modelos

Una vez generados ambos modelos con el mismo repertorio fonético procedemos a experimentar y evaluar de manera informal la efectividad del método. Para ello tomamos el modelo generado con *loc1_pal* y lo interpolamos con *CMU-ARCTIC-SLT* con diferentes pesos entre ellos. Esta interpolación, a cargo de HTS_engine, consiste a grandes rasgos en tomar ambos HMMs e interpolar sus características fonéticas y prosódicas para obtener un nuevo modelo. Para una explicación más detallada del tema ver la Sección 2.5.4

Para grados cercanos al 90 % de castellano + 10 % de inglés obtenemos los resultados esperables: las oraciones sintetizadas tienen un marcado acento castellano. Asimismo, en el otro extremo, 10 % de castellano + 90 % de inglés, la voz sintetizada, al igual que lo que se describió en la sección anterior, presenta problemas fonéticos graves, haciendo que las oraciones resulten poco naturales y difíciles o imposibles de comprender. Para grados intermedios de interpolación, 70 % de castellano + 30 % de inglés y 40 % de castellano + 60 % de inglés, observamos resultados más cercanos a los esperados, pudiendo apreciar en

las oraciones sintetizadas detalles distintivos como el fono [ɾ] más suavizado, o la pronunciación de vocales más abiertas, pero aún conservando cierto grado de inteligibilidad. De esta misma manera también pudimos apreciar en las oraciones sintetizadas cierta prosodia no familiar que también podría ser adjudicada a un hablante extranjero.

Como un efecto colateral de la interpolación, pudimos apreciar que cuanto más cercano es el grado de interpolación al modelo entrenado con *loc1_pal*, las características fonéticas más se asemejan a las de las oraciones del corpus *loc1_pal*. De manera análoga, cuanto más porcentaje de inglés tiene la mezcla, la voz más se asemeja a *CMU-ARCTIC-SLT*. Si bien esto no es un defecto importante, con el motivo de cambiar el menor número de variables en la experimentación sería deseable que la voz no presentara distintas características para distintos grados de interpolación.

Como una posible solución surge la posibilidad de utilizar Speaker-Adaptive Training sobre uno de los modelos, que describiremos a continuación.

## 2.4. Speaker-Adaptive Training

El Speaker-Adaptive Training es una técnica que permite tomar un modelo ya entrenado y adaptarlo para asimilar características de un nuevo hablante. Esta técnica nace de la idea de que construir un corpus de datos es costoso en espacio de almacenamiento, tiempo de grabación y etiquetado, por lo que resulta más económico generar una modelo base a partir de un gran corpus de datos y luego adaptarlo con características particulares del hablante especifico que queramos.

Nuestro objetivo para este trabajo es utilizar esta herramienta para aproximar las características de uno de los hablantes al otro para que sus identidades sean indistinguibles. Como prueba de concepto se entrenó el modelo principal con *CMU-ARCTIC-SLT* y se le realizó Speaker-Adaptive Training junto con *loc1_pal* utilizando los scripts provistos en la sección de descargas de HTS [12].

Dentro del adaptive training existen varias técnicas, en este trabajo utilizaremos *offline supervised adaptation*, que tiene como requisito adicional conocer las transcripciones fonéticas del segundo corpus.

Las pruebas no resultaron concluyentes. Las oraciones sintetizadas no solamente adquirían la identidad del segundo hablante sino también sus características fonéticas. Dicho de otra manera, si lo que buscábamos era obtener un hablante inglés que pudiera ser reconocido como el mismo locutor que *loc1_pal* pero con sus características fonéticas intactas (una pronunciación suavizada del fono [r] (*perro*), por ejemplo), lo que en realidad obtuvimos fue una voz idéntica a *loc1_pal*. Si bien existen indicios que indican que es posible generar un modelo con las características deseadas [10] [13], dada la complejidad del método y los largos periodos de entrenamiento (36 horas aproximadamente) decidimos abandonar este camino y continuar con la fase de evaluación perceptual sobre las voces generadas con las técnicas de interpolación.

## 2.5. Herramientas

En esta sección presentamos las herramientas y los comandos con los que se realizó el preentrenamiento, el entrenamiento y la generación de oraciones para la tesis.

### 2.5.1. Festival y Festvox: Generación de transcripciones fonéticas

Festival [17] es un framework que permite sintetizar habla. Además posee una gran variedad de APIs para el procesamiento de audios y generación de nuevos sistemas TTS. Festvox [18] a su vez expande sobre Festival, agregando todavía más herramientas relacionadas a la síntesis y generación de modelos, que van desde la generación de modelos prosódicos, hasta etiquetado automático de corpus.

Para este trabajo utilizamos Festival y Festvox para generar las oraciones requeridas tanto para el entrenamiento como para la síntesis de audios. Estas transcripciones consisten en una lista de fonos dividida en segmentos temporales y datos contextuales tales como la cantidad de sílabas en la palabra siendo transcrita, fonos que preceden y proceden al actual, etc. A modo de guía, a continuación mostraremos cómo utilizamos estas herramientas para generar las transcripciones fonéticas deseadas usando EHMM alignment.

Primero tenemos que generar un archivo *txt.done.data* donde estén los nombres de cada archivo de audio y su transcripción grafémica. Por ejemplo, en el siguiente recuadro podemos ver un extracto del archivo generado para SECYT_mm utilizado para este proceso:

```
( SECYT_mm_1_335 “Algunos dicen gamba en vez de pierna” )
( SECYT_mm_1_29 “El conjunto de las escenas se reitera en el galpón” )
( SECYT_mm_1_361 “Lluvia con truenos en Medellín” )
( SECYT_mm_1_619 “Rendían pleistecía vikingo conquistador” )
( SECYT_mm_1_110 “Llueve sobre las piedras de la pared” )
( SECYT_mm_1_102 “Las etapas del desarrollo infantil difieren según el niño” )
```

En este trabajo utilizamos Festival 2.4 [17], Festvox 2.7 [18] y speech_tools 2.4 [19] para la generación de transcripciones fonéticas. Para poder utilizarlos agregamos las siguientes variables de entorno en nuestro PATH:

```
export PATH=/project/festival/bin:$PATH
export PATH=/project/speech_tools/bin:$PATH
export FESTVOXDIR=/project/festvox
export ESTDIR=/project/speech_tools
```

Luego generamos los directorios, scripts y archivos necesarios para generar una nueva voz:

```
$FESTVOXDIR/src/clustergen/setup_cg uba es SECYT_mm
```

En la nueva estructura de archivos generada, copiamos los audios en la carpeta wav/ y el archivo *txt.done.data* previamente generado en la carpeta etc/.

Además en los archivos

```
festvox/uba_es__cg.scm
festvox/uba_es__clunits.scm
```

es necesario cambiar las dependencias

```
(require 'uba_es__phoneset)
(require 'uba_es__lexicon)
```

que contienen los símbolos fonéticos del español de España, por estas otras:

```
(require 'uba_es__phoneset_mex)
(require 'uba_es__lexicon_mex)
```

que contienen el conjunto de símbolos fonéticos del español mexicano, que se aproximan mucho a los del castellano rioplatense (no contiene [th], por ejemplo).

Finalmente, corriendo los siguientes comandos:

```
./bin/do_build build_prompts
./bin/do_build label
./bin/do_build build_utts
```

se realizará el proceso de alineamiento y transcripción automática. Una vez finalizado se habrán generado las transcripciones fonéticas con formato .utt en el directorio festival/utts, que entre otra metadata tiene codificados los fonos de la oración, sus principios y sus finales.

De manera análoga, esta herramienta permite crear partituras utilizadas para la síntesis. Simplemente generando un archivo *txt.done.data* con oraciones que se quieran sintetizar, y corriendo el script

```
./bin/do_build build_prompts ./synth/txt.done.data
```

En la carpeta utt gen/prompt-utt se habrán generado los .utt necesarios para la síntesis.

### 2.5.2. HTS

HTS [6] es un framework de entrenamiento y síntesis de sistemas TTS basado en HMMs que modela simultáneamente la duración, el espectro (mel-cepstrum) y la frecuencia principal ($f0$) utilizando una combinación de HMMs.

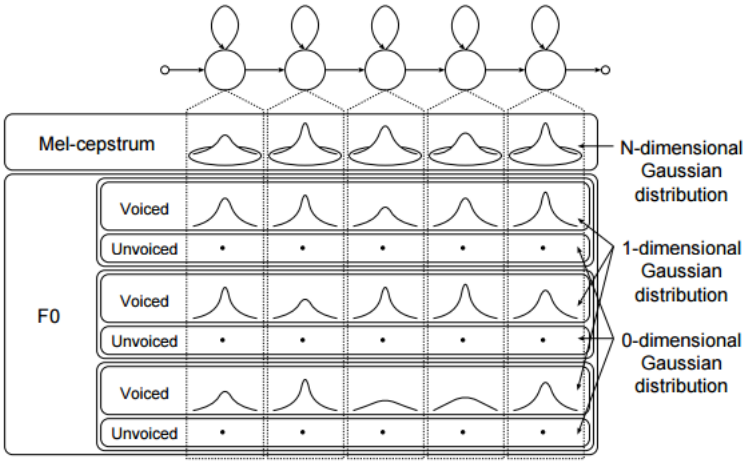



Figure 5.3: structure of HMM.

*Fig. 2.1:* Estructura de un HMM (tomado de [11], página 41)

La Figura 2.1 resume la estructura de un HMM usado para síntesis del habla. El espectro y la frecuencia fundamental son modelados en paralelo usando vectores separados. En particular el espectro es modelado como un vector de Gaussianas $n$ dimensional, mientras que la frecuencia principal es modelada como un conjunto de vectores de Gaussianas de dimensión uno y cero.

Al mismo tiempo HTS toma la decisión de modelar la información prosódica dentro de este mismo framework. Para esto, las distribuciones para el espectro, la frecuencia principal y las duraciones son clusterizadas independientemente utilizando la información contextual extraída de los audios de entrenamiento. A modo ilustrativo en la Figura 2.2 se muestra una esquematización de un HMM resultante, utilizando árboles de decisión para clusterizar los datos. Notar que cada hoja del árbol resultante coincide con un vector $n$ dimensional de Gaussianas o un conjunto de vector de Gaussianas de dimensión cero y uno, según corresponda al espectro o a la frecuencia principal.

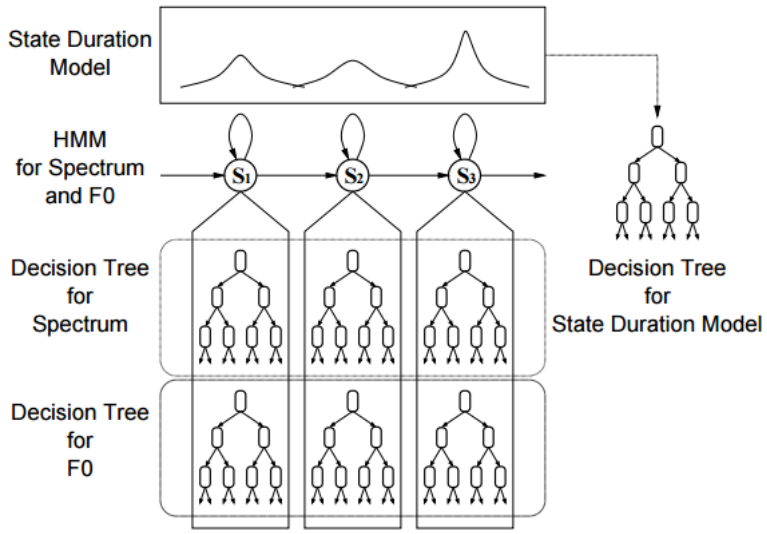



Figure 5.4: Decision trees.

*Fig. 2.2:* Esquema HMM generado utilizando árboles de decisión (tomado de [11], página 45)

Si bien existen muchas maneras de clusterizar el conjunto de fonos, que pueden variar desde algoritmos simples hasta técnicas que utilicen redes neuronales, para este trabajo todos los entrenamientos y clusterizaciones de datos se realizaron con árboles de decisión.

Por otro lado, como información contextual para el entrenamiento se tomaron los dos fonos por anteriores y posteriores a cada fono y la siguiente información fonética.

- Modo de articulación del fono:
  - 
- Punto de articulación del fono:
  - 
- La perspectiva articulatoria (anterior, central o posterior).
- Si el fono es una vocal o una consonante.
- En caso de ser una vocal, a que categoría pertenece: por ejemplo para el fono [$i$] :$i$ (no acentuada), $i0$ (diptongo), $i1$ (acentuada).
- En caso de ser una vocal, su redondeamiento vocálico.
- En caso de ser una consonante, si es débil o fuerte.

De esta manera HTS espera tener una voz más dinámica, que para diferentes valores contextuales darán diferentes modelos acústicos para cada fono.

En la Figura 2.3 se muestra el resultado de un fragmento de uno de los árboles de decisión generado para modelar la duración de un fono. En base a este modelo, el sistema podrá inferir, por ejemplo, que si el fono actual no es nasal (C-Nasal) seguido de un stop

(R-Stop), que no es el fono *l* estará modelado por función de probabilidad Gaussiana definida en *dur_s2_7*.

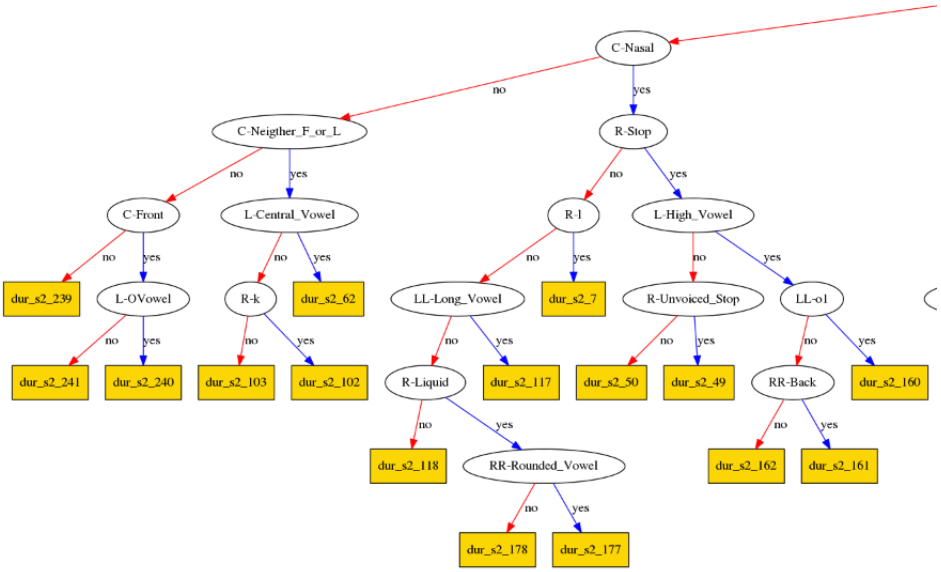



*Fig. 2.3:* Árbol de decisión generado a partir de los datos para la duración de un HMM

En las primeras iteraciones del desarrollo no contábamos con la información acústica, por lo que se generaron modelos carentes de información contextual. En estos primeros modelos se pudo apreciar una calidad muy inferior en los audios generados, sonando estos sumamente metálicos y carentes de prosodia. Esto se debía, posiblemente, a que los árboles de decisión no tenían información contextual suficiente para ser construidos de manera efectiva, resultando en una mala generalización y malos audios sintetizados. Tras agregar los factores contextuales y realizar algunas pruebas de concepto, pudimos comprobar que las voces sonaban con mucha mejor calidad.

### 2.5.3. Entrenamiento

Desde un punto de vista puramente técnico, utilizar HTS para entrenar un modelo es bastante sencillo. Asumiendo que todos los paquetes necesarios fueron instalados, es posible entrenar una nueva voz adaptando los scripts disponibles en la página de descargas de HTS.

Para ello, reemplazamos los audios en la carpeta data/raw con aquellos que se quieran utilizar y los utts correspondientes previamente generados. Además, como adelantamos en la sección anterior, le indicamos a HTS qué información contextual se utilizará para la clusterización, por lo que es necesario modificar el archivo data/questions/questions_qst001.hed con la información contextual apropiada para una voz en castellano. En el apéndice **??** se presenta el archivo utilizado para Loc1_pal.

Una vez finalizadas estas modificaciones, en la carpeta data/ de la demo puede iniciarse el pre-entrenamiento de la siguiente manera:

```
make
```

Esto extraerá features acústicos del audio y construirá los archivos de entrenamiento, entre otras cosas. Finalmente para dar comienzo al entrenamiento, ejecutamos en la carpeta raíz:

```
perl scripts/Training.pl scripts/Config.pm > train.log 2> err.log
```

Una vez que se completa el entrenamiento, podemos encontrar en la carpeta voices/qst001/ver1 el modelo generado (.htsvoice).

Para este trabajo todos los audios usaron sampling rate de 48kHz, precisión de 16bits, mono. Además HTS requiere que explicitemos un rango de extracción para la frecuencia fundamental de la voz. Tanto para *SECYT-mujer* como para *Loc1-Pal* el rango utilizado utilizado fue desde 100hz hasta 350hz, mientras que para *CMU-ARCTIC* el rango de extracción fue desde 110kHz hasta 280kHz.

Estos parámetros y muchos otros pueden ser configurados fácilmente corriendo

```
./configure
```

en la carpeta raíz del proyecto. En la próxima sección detallaremos como utilizar varios .htsvoice generados para mezclar y sintetizar una nueva voz.

### 2.5.4. HTS_engine

Finalmente, para generar voces con acento extranjero se utilizó hts_engine. Esta herramienta, entre otras cosas, permite interpolar entre varios modelos, para producir un nuevo modelo con una mezcla de la carga fonética y prosódica. Esto nos brinda un gran rango exploratorio y nos permite ajustar la carga fonética de los modelos entrenados previamente para cumplir con nuestro objetivo de sintetizar habla en castellano con acento inglés.

A grandes rasgos, la interpolación consiste en tomar los vectores generados anteriormente durante el entrenamiento e interpolar sus funciones de densidad Gaussianas para obtener una nueva. En la Figura 2.4 (extraída del trabajo [16]) puede verse la interpolación de $N$ HMMs, cada uno con peso arbitrario $a_1$, $a_2$, ..., $a_N$, que generan un nuevo modelo $\Lambda$.

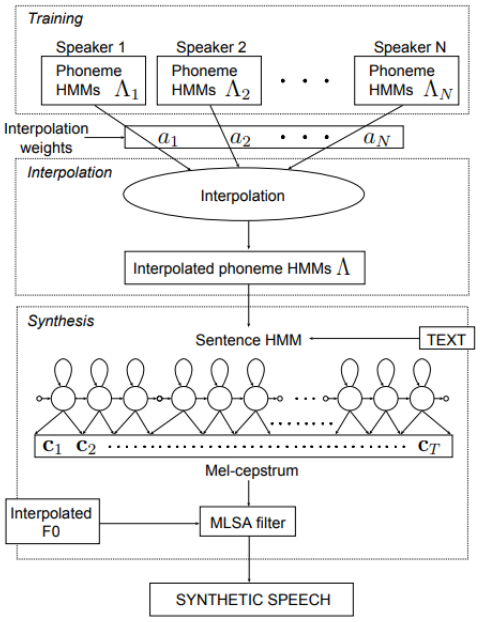



*Fig. 2.4:* Block diagram of speech synthesis system with speaker interpolation. (tomado de [11], página 70)

Una vez obtenido el nuevo HMM, el proceso de síntesis puede ocurrir como para cualquier otro modelo, ilustrado en la etapa de síntesis de la figura. El procedimiento para sintetizar una nueva oración es bastante simple. Asumiendo que todas las dependencias fueron instaladas de manera correcta, con el siguiente comando es posible utilizar los modelos generados en *cmu_us_arctic_slt.htsvoice* y *models/loc1_pal.htsvoice* para interpolar con peso 0,7 y 0,3 respectivamente, generar un nuevo modelo y sintetizar la oración presente en el archivo in.lab.

```
hts_engine -m models/cmu_us_arctic_slt.htsvoice -m models/loc1_pal.htsvoice -i 2
0.7 0.3 -ow out.wav in.lab
```

Esta herramienta permite además modificar el pitch, la duración del audio y otros aspectos de la síntesis. La documentación completa puede encontrarse en la pagina web: http://hts-engine.sourceforge.net/.

# 3. EVALUACIÓN PERCEPTUAL

En esta sección intentamos validar que el habla sintetizada por los modelos generados realmente puede ser identificada como perteneciente a personas de habla inglesa, y al mismo tiempo evaluar sus grados de inteligibilidad. Para eso se condujo una encuesta perceptual donde a cada participante se le presentó una oración sintetizada con distintos grados de mezcla de español e inglés y se le pidió que la transcribiera y que intentara identificar la nacionalidad del hablante.

La encuesta se realizó a través de Internet, con el mismo set de instrucciones para todos los participantes y pidiendo como requisito la utilización de auriculares. Cada participante podía contestar un máximo de 5 veces, presentándoles siempre oraciones distintas.

Nos propusimos como objetivo conseguir 5 respuestas para cada uno de los 50 audios sintetizados, momento en el cual se cerró la posibilidad de contestar. Se llevó a cabo desde el 18 de octubre de 2017 hasta el primero de diciembre del mismo año, tiempo durante el cual fue publicitada en distintas redes sociales y listas de emails de la facultad.

Con el objetivo de no influir en las respuestas de los participantes, se procuró darles la información mínima indispensable para completar la encuesta. Por este motivo, en ningún momento de la encuesta se especifica el objetivo real de la misma. Con la intención de homogeneizar la muestra, fue requisito obligatorio utilizar auriculares para la encuesta. También se le pidió a cada participante que la realizara en un lugar silencioso y tranquilo.

A continuación, describimos la forma en que se construyeron los audios usados en esta evaluación y la interfaz de la encuesta. Luego mostramos los datos demográficos obtenidos. Más adelante, continuamos con un análisis exhaustivo de inteligibilidad y origen atribuido a las oraciones. Por último, para evaluar la hipótesis original, compondremos estos dos ejes para dilucidar el grado de validez de los resultados.

## 3.1. Audios sintetizados

Para evitar que el participante pudiera deducir las palabras de los audios a partir de las palabras vecinas, las mismas fueron generadas de manera *semánticamente impredecible*. Esto significa que a partir de una lista de sustantivos, adjetivos, determinantes y verbos se generaron oraciones de manera aleatoria con la siguiente estructura:

*Determinante Adjetivo Sustantivo Verbo Determinante Sustantivo*

Luego, para asegurarnos de estar cubriendo todos los posibles fonos del castellano, las oraciones fueron modificadas para ser fonéticamente balanceadas. Esto significa que incluimos entre cinco y diez veces cada fono perteneciente a una consonante (del castellano) y al menos veinte veces cada fono perteneciente a una vocal.

Los oraciones generadas fueron:

- Oración 1: Mi montaña aguileña recorrió la esquina
- Oración 2: Aquel fuerte vidrio prefirió aquel botón
- Oración 3: Este enjoyado juez comprará nuestro corchete

- Oración 4: Tu estrecho posavasos gritó la fechoría
- Oración 5: Nuestro nublado tigre concluyó a este chupetín
- Oración 6: Su profundo riñón apoyó a Julio
- Oración 7: El frío churrasco oyó lo de Polonia
- Oración 8: Las acongojadas cotorras sonrieron a mi círculo
- Oración 9: Ese gruñón perro prometió a esos cuñados
- Oración 10: El nudillo argentino perdió su vaso

Para cada una de estas diez oraciones se varió el nivel de mezcla entre 30 % de inglés + 70 % castellano hasta 70 % de inglés + 30 % de castellano, con 10 % de incremento. De esta manera, para cada oración hay 5 mezclas diferentes, lo que hace un total de 50 audios sintetizados diferentes.

## 3.2. Interfaz

En esta sección se presenta la interfaz utilizada para realizar la encuesta junto con las decisiones de diseño más relevantes. En la Figura 3.1 se presenta la página con la que todos los participantes fueron recibidos. A fin de conocer de manera general la demografía encuestada, a cada participante se le pidió que indicara el rango correspondiente a su edad, yendo desde 18 a 25, 26 a 35, y así de diez en diez.

Se le pidió, además, que indicara su género: “masculino”, “femenino”, “otro”, “no contesta” y la provincia donde pasó la mayor parte de sus primeros diez años de vida. Consideramos que estos datos son importantes para el estudio ya que dependiendo de ellos los resultados variarán indefectiblemente. Por ejemplo la transcripción que obtendremos de un participante de 50 años de Capital Federal será distinta a la de alguien de 18 años de Córdoba. El diferente uso de los alofonos, modismos y variantes prosódicas y capacidades auditivas jugarán un papel importante en la interpretación de la oración y la apreciación del origen del hablante.


*Fig. 3.1:* Primera pantalla de la encuesta, en la cual se recaban datos personales

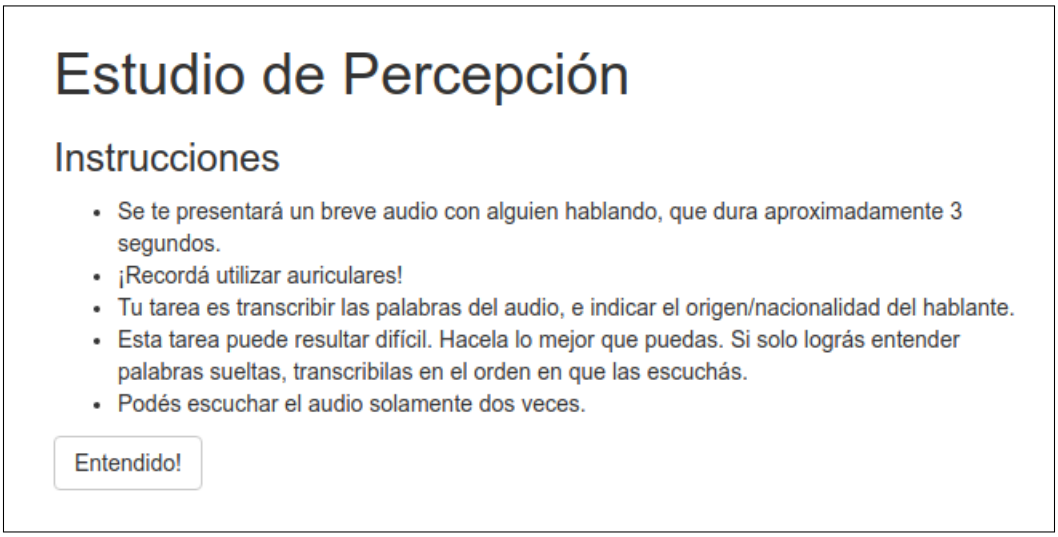



*Fig. 3.2:* Pantalla con las instrucciones de la encuesta

Como puede verse en la Figura 3.2, una vez completados estos datos, a cada participante se le presentó otra vista con las instrucciones específicas para completar la encuesta. Una vez presionado el botón de "Entendido!" se les presentó el primer audio, que podían escuchar un máximo de 2 veces, una caja de texto libre donde ingresar la transcripción del

mismo y una caja de texto libre donde podían escribir cual consideraban que era el origen de la nacionalidad correspondiente a la voz, como puede apreciarse en la Figura 3.3.

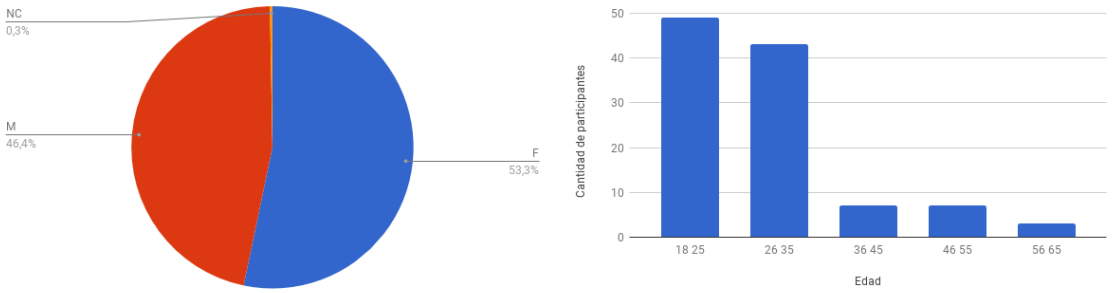

*Fig. 3.3:* Pantalla de transcripción

Una vez guardada la respuesta, se le preguntó si quería continuar transcribiendo otro audio. En caso de haber completado cinco audios, se le mostró un mensaje indicando que ya podía cerrar la encuesta:

> ¡Muchas gracias por tu participación! Ya podés cerrar la ventana del navegador.

## 3.3. Datos demográficos

Se encuestaron 109 participantes de los cuales se obtuvieron 352 resultados. Como puede verse en la Figura 3.4a del total de participantes, 49 pertenecían al rango comprendido entre 18 y 25 años, 43 estaban en el rango 26-35. 17 de los participantes eran mayores a 35 años.




*(a)* Distribución de géneros *(b)* Cantidad de participantes por rango de edad

*Fig. 3.4:* Datos demográficos de los participantes

Con respecto al género de los participantes (Figura 3.4), 187 respuestas fueron brindadas por participantes del género femenino mientras que 163 respuestas fueron brindadas por participante del género masculino .

Con respecto de la región en que cada participante pasó su infancia puede verse en la Figura 3.5 una predominancia de personas del Gran Buenos Aires con 45 %, seguido por un 30 % que pasaron su infancia en Capital Federal. Menos del 25 % pertenece al resto de las provincias Argentinas. Además, 10 personas contestaron que se criaron fuera del país.

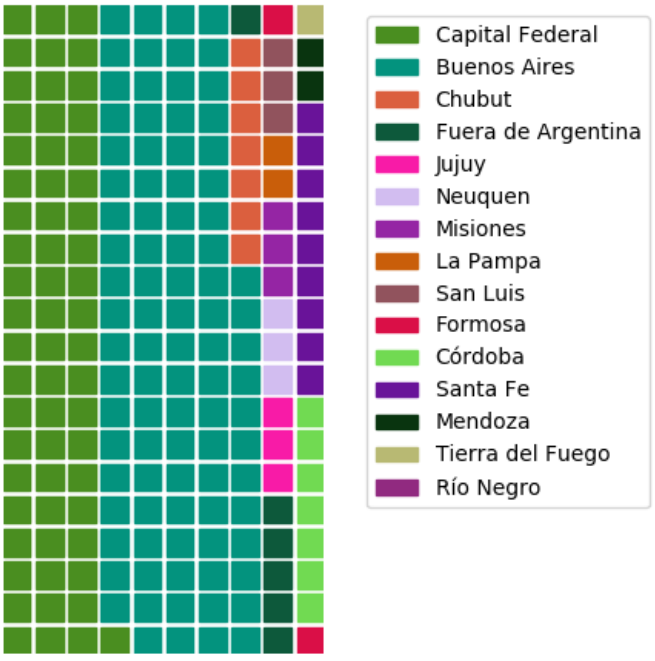



*Fig. 3.5:* Distribución Territorial

## 3.4. Análisis de la inteligibilidad

A continuación intentaremos medir la inteligibilidad de cada una de las oraciones en base a las respuestas obtenidas en la encuesta. Para ello tomaremos la oración transcripta por los participantes y mediremos cuán lejos o cerca está de la oración original.

Para esto utilizaremos un concepto denominado *distancia de Levenshtein*. La distancia de Levenshtein consiste en calcular la menor cantidad posible de inserciones, remociones o reemplazos de caracteres que son requeridos para transformar una oración en otra. Así por ejemplo, para transformar *cosa* en *cal* se requieren 3 transformaciones: reemplazar *o* por *a* final, reemplazar *s* por *l* y remover la *a*. Por lo tanto, la distancia de Levenshtein entre estas dos palabras es de 3. Además, se considerará un reemplazo cualquier acento, por lo que *á* y *a* tendrán distancia 1 pero no así el reemplazo de mayúsculas y minúsculas, por lo que *a* y *A* tendrán distancia 0.

En la Figura 3.6 presentamos los resultados generales obtenidos sin ningún tipo de modificación a las transcripciones ingresadas por los sujetos. En el eje $x$ presentamos los grados de interpolación de inglés-castellano, que varía entre 30 % inglés + 70 % castellano (desde ahora, 30 % inglés) y 70 % inglés + 30 % castellano (desde ahora, 70 % inglés). En el eje $y$ presentamos la distancia de Levenshtein entre la oración objetivo y aquella transcripta por cada participante. Cada uno de los boxplots describe la distancia mínima obtenida y la distancia máxima (los bigotes), como así también el primer y tercer cuartil (el piso y el techo de la caja) y la mediana (línea interior que atraviesa la caja). Adicionalmente podemos observar con círculos vacíos los outliers de la muestra.

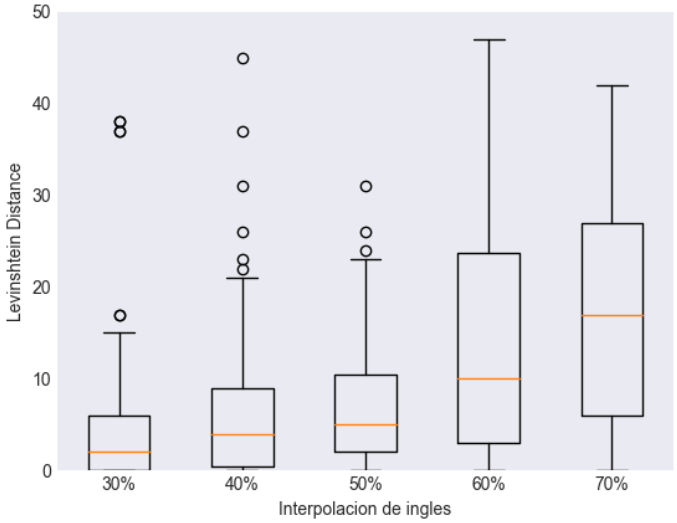



*Fig. 3.6:* Distancia de Levenshtein para distintos grados de interpolación

Como puede observarse hasta el 50 % de inglés en la interpolación, la distancia entre el primer y tercer cuartil es menor a 10 caracteres, siendo la mediana de 5. Pasados el 60 % de inglés, se observa un aumento brusco en la distancia intercuartiles, la distancia entre el primer y tercer cuartil pasa a ser cercana a 20 caracteres, y la mediana 10 caracteres en el caso de 60 % inglés y 15 en el caso del 70 %. En las próximas secciones intentaremos encontrar una explicación intuitiva a estos números.

## 3.5. Problemas en las transcripciones y normalización

Analizando detenidamente las transcripciones obtenidas podemos observar algunas fallas sistemáticas que podrían generar ruido en el análisis. Por ejemplo algunos de los participantes escribieron de manera diferente las secciones de la oración que no comprendieron. Por citar algunos ejemplos, muchos de ellos escribieron: "..." o simplemente omitieron la palabra, mientras que una minoría escribió cosas como "***", "???", "*blablabla*". En los casos donde el participante no comprendió ningún segmento de la oración, es común observar expresiones como "*no entendí nada*", "*nada*", etc.

Por otra parte, es común la utilización innecesaria de signos de puntuación. Estos varían desde puntos finales para expresar el final de la oración hasta expresiones de confusión tales como "(?)". En un caso, un participante transcribió *"tu estrecho posavasos", grito la fechoría*, cuando la oración original solo decía *tu estrecho portavasos gritó la fechoría*. También pueden verse omisiones de acentos y faltas ortográficas en palabras que no presentan ambigüedades, como por ejemplo: "*grunion*" en vez de "*gruñón*".

Todas estas expresiones, modismos, y faltas ortográficas tienen un efecto directo y negativo sobre lo que estamos intentando medir. Así, por ejemplo, un participante que haya transcripto "*no entendí nada*" en su respuesta devolverá una distancia de Levenshtein distinta de aquel que simplemente dejó el campo vacío, cuando en realidad estas dos respuestas expresan el mismo grado de comprensión del texto.

Con el objetivo de reducir la variabilidad en la muestra, decidimos realizar una limpieza de los datos. Intentando mantener el espíritu de las transcripciones, buscamos uniformizarlas para poder realizar un mejor análisis. Así, consideramos que si un participante escribió "..." en medio de una oración, su intención era la darnos a entender que no comprendió parte de la misma. Para nuestro análisis, esto sería equivalente a no haber escrito nada. Bajo este razonamiento, consideramos que los siguientes cambios no presentan alteraciones graves en las respuestas de los participantes:

- Corrección de "ni" por "ñ" en la palabra *grunion*.
- Remoción de todos los signos de puntuación: comas, puntos, "(?)"
- Reemplazo de oraciones como *blabla*, *no entendí* o cualquier otra expresión que indique ininteligibilidad de una palabra u oración por la cadena vacía.
- Corrección de acentos en palabras no ambiguas: *botón*, *prefirió*, *recorrió*, *chupetín*, *riñón*, *gruñón*.

Aquellas palabras que presentan ambigüedad, como *concluyo*, no fueron modificadas, ya que tanto *concluyó* como *concluyo* son válidas. El participante podría haber interpretado la palabra con cualquiera de las dos connotaciones, cambiando el significado de la interpretación y su distancia de Levenshtein.

Confiamos en que esta limpieza nos ayudaría a disminuir el error de los resultados y también, nos permitiría interpretar de manera más intuitiva el significado de la distancia de Levenshtein en cada caso.

## 3.6. Análisis de la inteligibilidad con los datos normalizados

Al igual que en la Sección 3.4, en la Figura 3.7 presentamos los resultados de la muestra, luego de la normalización.

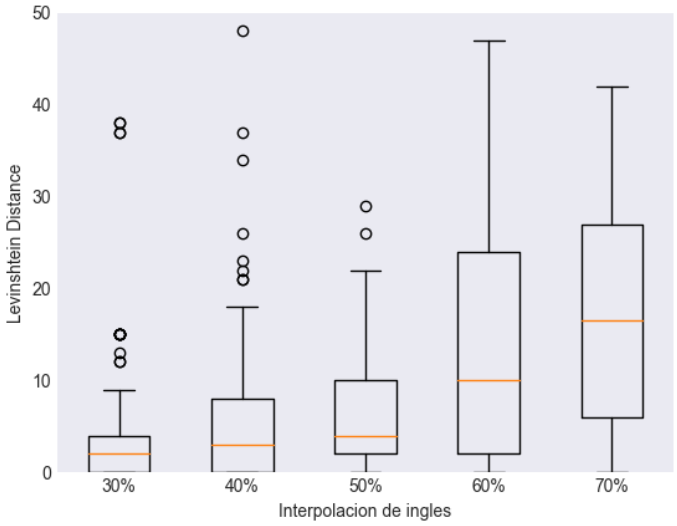



*Fig. 3.7:* Distancia de Levenshtein para distintos grados de interpolación con datos normalizados

En esta figura podemos observar que para una interpolación de inglés de 30 %, 96 de los 106 participantes obtuvieron una distancia menor a los 10 caracteres en la transcripción del texto, mientras que los 10 restantes, una distancia mayor a los 10 caracteres. Podemos ver además que la distancia entre el primer y tercer cuartil es menor a 5 caracteres con una mediana cercana a 2

Para la mezcla con 40 % de interpolación de inglés, de un total de 67 participantes, 57 anotaron una distancia de Levenshtein menor a diez caracteres, 7 una distancia entre 10 y 30 caracteres y 3 una distancia mayor a 30. La distancia intercuartil es de aproximadamente diez caracteres, con una mediana muy similar a la mezcla 30 % inglés.

Para la interpolación 50 % de inglés, de un total de 75 participantes, 57 lograron transcribir el audio con una distancia menor a 10 caracteres, mientras que 14 anotaron una distancia entre 10 y 20 caracteres y 4 una distancia mayor a 20. De manera similar que para 40 % la distancia intercuartil es de aproximadamente diez caracteres y la mediana de 3 caracteres.

### 3.6.1. Análisis estadístico

En la siguiente sección, realizaremos un análisis estadístico con el objetivo de vislumbrar cuan estadísticamente significativas son las diferencias en inteligibilidad entre los saltos de interpolación.

En primera instancia evaluamos usar el Student's two-sample t-test para este análisis. Este test pide como requerimiento que las muestras tengan una distribución normal. Para probar esto aplicamos el test de Shapiro sobre los datos normalizados.

| Interpolación | p-valor | W |
|---|---|---|
| 30 % | 6.875e-13 | 0.5832 |
| 40 % | 6.875e-13 | 0.6950 |
| 50 % | 1.271e-06 | 0.8674 |
| 60 % | 2.932e-06 | 0.8694 |
| 60 % | 0.00215 | 0.9380 |

*Tab. 3.1:* Test de Shapiro sobre los datos normalizados

Como puede verse en la Tabla 3.1 los 5 tests dan $p < 0{,}01$, entonces rechazamos la hipótesis nula de normalidad, y concluimos que los datos no siguen una distribución normal. Descartamos la posibilidad de usar el Student's two-sample t-test y procedemos a utilizar el test no paramétrico Mann-Whitney-Wilcoxon, con restricciones más laxas. Este test pide como único requerimiento que las observaciones de ambos grupos sean independientes. Cabe recordar que, por diseño, nuestras observaciones cumplen con este requerimiento, ya que el sistema siempre asignaba audios con diferentes oraciones a un mismo participante, por lo que aunque contestaran más de una vez, las respuestas eran independientes entre sí.

| Interpolaciones | p-valor | W |
|---|---|---|
| 30 % vs. 40 % | 0.1367 | 2041.5 |
| 40 % vs. 50 % | 0.1148 | 2132 |
| 50 % vs. 60 % | 0.0051 | 1921.5 |
| 60 % vs. 70 % | 0.0708 | 1956 |

*Tab. 3.2:* Resultados test no paramétrico Mann-Whitney-Wilcoxon

Como se ve en la Tabla 3.2 la diferencia en distancia de Levinshtein entre 30 % y 40 % de interpolación de inglés no es estadísticamente significativa ($p = 0{,}1367$). Así tampoco lo es para el salto de 40 % a 50 % de interpolación de inglés ($p = 0{,}1148$). En cambio, al pasar de 50 % a 60 % la diferencia es significativa ($p = 0{,}0051$), lo cual se interpreta como que para este salto la inteligibilidad se deteriora más sensiblemente que para los saltos anteriores. Por último, de 60 % a 70 % la diferencia de distancias de Levinshtein es aproximadamente significativa ($p < 0{,}1$).

### 3.6.2. Análisis por oración

Tras aplicar la normalización de los datos descripta en la Sección 3.5, tratamos ahora de encontrar una interpretación intuitiva para las distancias de Levenshtein obtenidas. Por ejemplo, tomando la oración 8 de las frases utilizadas en la experimentación:

- "Las acongojadas cotorras sonrieron a mi círculo"

Podemos observar las siguientes transcripciones extraídas de los resultados:

- Distancia 0: "Las acongojadas cotorras sonrieron a mi círculo"
- Distancia 10: "Las acontojadas culturas sonrieron en semicírculo"

- Distancia 20: “Plaza sombreada con sombrero sonrieron en mi círculo”
- Distancia 30: “sonrieron en mi círculo”
- Distancia 40: “círculo”
- Distancia 48: “”

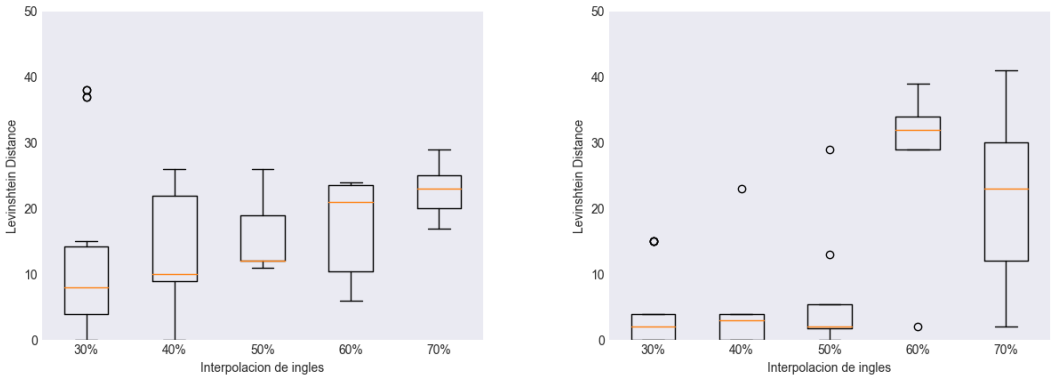
Para todas las interpolaciones enunciadas previamente, los errores más comunes varían desde falta de acentos en palabras como “concluyó” hasta faltas de inteligibilidad en palabras con cierta complejidad fonética como “aguileña” o “gruñón”. En el Apéndice 5.1 listamos todas las transcripciones ingresadas por los participantes del estudio.

Como caso particular, en la oración 3 (Figura 3.8a), “este enjoyado juez comprará nuestro corchete”, observamos que la mayoría de los participantes cometieron errores al transcribir la palabra “juez”, que confundieron de manera sistemática con palabras sonoramente similares como “fue”, “enjoyado”, que transcribieron como “enfollado” o “enrollado”, y la conjugación del verbo “comprar” que transcribieron como “comprando” o “comprar”.

Buscando una explicación a estos errores y revisando el mapeo de fonos realizado previamente, descubrimos que el mapeo de [hh] a [g] es erróneo. Esto produjo que palabras como “gato” sonaran más como “jato” ([hh] [a] [t] [o]). Adicionalmente descubrimos que ningún fono del inglés estaba siendo mapeado al fono [x], correspondiente al grafema j en castellano. Esto produjo que cuando el sintetizador interpola entre el fono del castellano y el fono inexistente del inglés (que HTS toma como ruido blanco), el resultado fuera una mezcla entre ruido y [x]. Al ser la [x] ruido blanco (consonante fricativa), no pudimos reconocer el error en las instancias preliminares de evaluación. Esto explicaría por qué algunos participantes tuvieron problemas transcribiendo la palabra “juez”.

Continuando con el análisis, para la interpolación de 60 % inglés, podemos observar un aumento notable de la variabilidad en las respuestas. Para el primero, de las 70 respuestas obtenidas, 40 participantes transcribieron el audio con una distancia menor a 10 caracteres, 6 anotaron una distancia entre 10 y 20 caracteres, y 24 transcribieron el audio con distancia mayor a 20 caracteres. Además, podemos ver en la Figura 3.8a que la distancia entre el primer y tercer cuartil es muy marcada, siendo de 20 caracteres con una mediana igual a 10.

Consideramos que este salto tan marcado en la distancia intercuartil puede deberse a que existen características particulares de los participantes y sus capacidades para discernir palabras, incluso cuando presentan defectos en la pronunciación del hablante. Para reforzar esa hipótesis puede verse en la figura 3.8b como en la oración 4 para la interpolación de 70 % inglés, 2 participantes de los 9 que realizaron la transcripción, obtuvieron distancias 2 y 6 en sus transcripciones.




*(a)* Oración 3 *(b)* Oración 4

*Fig. 3.8:* Distancia de Levenshtein para distintos grados de interpolación con datos normalizados

Además, comparando la oración 10 (Figura 3.9) "El nudillo Argentino perdió su vaso" con las anteriores, podemos apreciar que parecería haber características particulares en las oraciones o en el modelo generado que afectan la comprensión de los audios. Para esta oración, con 70 % de inglés en la interpolación, 6 de los 8 participantes obtuvieron una transcripción del audio con distancia menor a 15, tanto que para la oración 3[1], con ese mismo grado de interpolación, todos los participantes transcribieron el audio con distancia de Levenshtein mayor a 15. Esto parecería indicar que o bien la dificultad de las oraciones es variable o, lo que es todavía más probable, llegado cierto punto en la interpolación, algunos fonos empiezan a "romperse" o se alejan demasiado del fono del castellano correcto y terminan por disminuir la claridad de la voz.

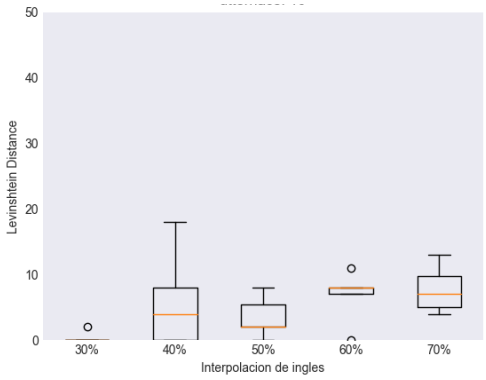



*Fig. 3.9:* Distancia de Levenshtein para distintos grados de interpolación con datos normalizados oración 10

En esta sección analizamos distintos problemas y particularidades en las oraciones que nos permitieron descubrir errores implementativos y teorizar sobre las características del modelo generado. Pudimos ver que la comprensión de una misma oración puede variar mucho dependiendo del oyente. También pudimos ver que, incluso utilizando el mismo

[1] Tener en cuenta que ambas oraciones se encuentran afectadas por el mismo mapeo incorrecto de la [x] como fue discutido previamente

grado de interpolación, no todas las oraciones presentan la misma dificultad a la hora de ser interpretadas.

## 3.7. Análisis del origen percibido

En esta sección analizaremos los resultados de los orígenes o nacionalidades que los participantes de la encuesta atribuyeron a las voces sintetizadas. Dado que en esta instancia se le permitió a los participantes ingresar texto libre, las respuestas resultaron bastante heterogéneas. Los participantes interpretaron la consigna de distintas maneras, pudiendo encontrarse respuestas que no pueden ser atribuidas a una nacionalidad. Como ejemplo de algunas respuestas pueden encontrarse cosas como: "Latino", "Anglo", "Robot", "España (sur)". Consideramos que las respuestas de la índole "robot", "es una voz artificial", no aportan información para esta investigación. Por esta razón, en esta instancia decidimos agrupar las respuestas en cuatro conjuntos:

- Hispanoparlante: "Latino", "Argentino", "Español", "Uruguayo", "Centroamericano", "Boliviano", " Mexicano","Colombiano".
- Angloparlante: "Estadounidense", "inglés", "Irlandés", "Canadiense", "Anglo".
- No sabe/No contesta: "Robot", "no se".
- Otro: "Ruso", "Brasiltiño" (sic).

Con estas agrupaciones, en la Figura 3.10 presentamos las nacionalidades atribuidas a la voz generada para cada punto de la interpolación.

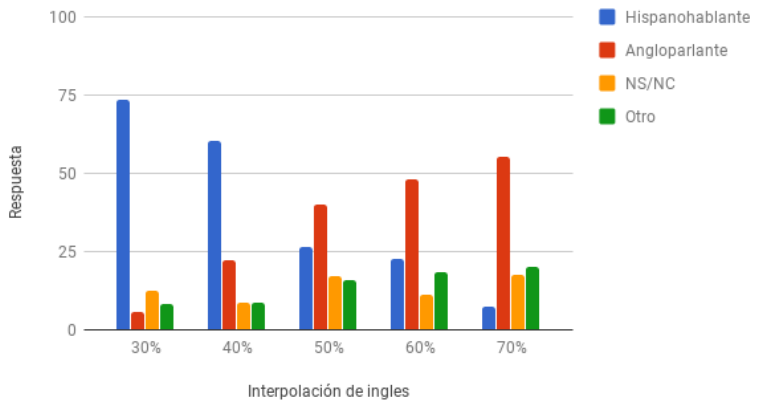



*Fig. 3.10:* Interpolación inglés vs respuesta de los participantes

De estos resultados podemos observar que con 30 % de interpolación de inglés, los participantes coinciden ampliamente en que la voz puede atribuirse a una persona de habla nativa española.

En la figura pueden verse dos tendencias muy marcadas. La primera es que, a medida que aumenta el grado de interpolación de inglés, disminuye de manera lineal la cantidad de

participantes que afirman que el hablante es hispánico. Dicho de otra manera, los datos parecen sugerir que con cada salto en la interpolación de inglés, la proporción de participantes que afirma que la voz pertenece a un Hispanohablante, disminuye en aproximadamente 7 puntos perceptuales.

Una tendencia opuesta puede notarse en los participantes que afirman que la voz es de un hablante anglosajón. A medida que aumenta el grado de interpolación de inglés, el porcentaje de atribuciones de la voz a un hablante Anglosajón aumenta aproximadamente en un 7 % cada vez.

Por otro lado, tanto para el conjunto "Otro" como para "NS/NC" podemos apreciar un leve aumento monótono entre los aumentos de interpolación de inglés.

### 3.7.1. Análisis estadístico

Para darle peso estadístico a las afirmaciones realizadas en el apartado anterior, intentaremos analizar si las tendencias descendentes/ascendentes de las proporciones de sujetos que percibieron a los audios como español/inglés responden a un relación lineal con respecto al nivel de interpolación de inglés. Para eso, buscaremos ajustar un modelo lineal sobre los datos.

Como primer análisis, ajustamos un modelo lineal para la atribución del audio a una persona de habla hispana. Nuestra variable independiente será el porcentaje de interpolación de inglés en el modelo y como variable dependiente tendremos el porcentaje de sujetos que percibieron el audio como un hablante español.

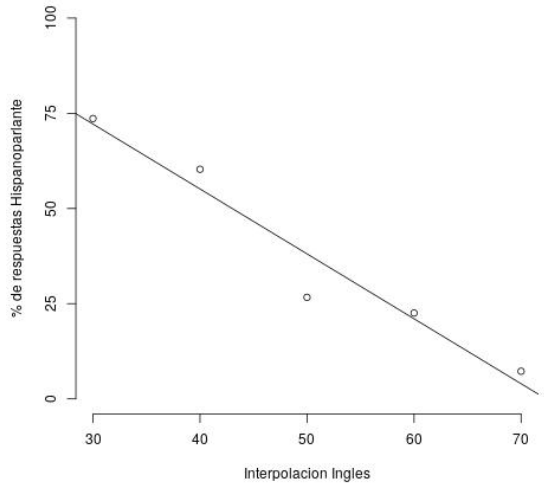



*Fig. 3.11:* Modelo lineal ajustado para la percepción de origen hispanoparlante

Como puede verse en la Figura 3.11 el modelo lineal ajustado explica satisfactoriamente los datos, con un nivel elevado de significancia estadística ($p < 0{,}01$).

Como segundo análisis, ajustamos otro modelo lineal para la percepción de atribuciones de la nacionalidad a una persona angloparlante. En este caso la variable independiente será nuevamente el grado de interpolación de inglés y como variable dependiente tendremos el porcentaje de sujetos que percibieron el audio como un hablante inglés.

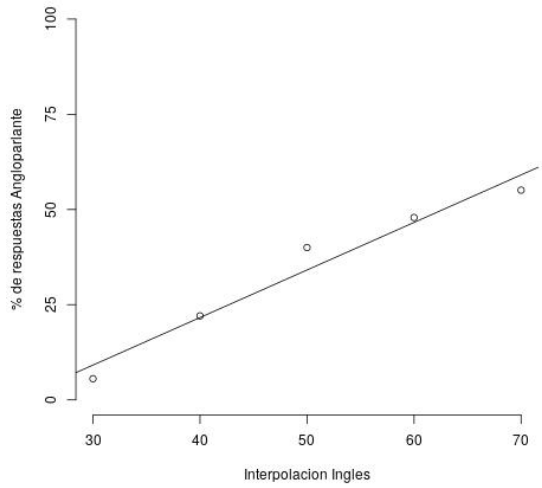



*Fig. 3.12:* Modelo lineal ajustado para la percepción de origen angloparlante

En la Figura 3.12 puede verse que el modelo lineal ajustado explica satisfactoriamente los datos, con un nivel elevado de significancia estadística ($p < 0{,}05$).

De esto podemos concluir que existe evidencia estadística que indica que aumentar el grado de interpolación resulta en más participantes reconociéndolo como una persona de habla inglesa y no como un hablante oriundo de un país de habla hispana.

### 3.7.2. Análisis por oración

Observando las oraciones una por una, podemos ver algunas discrepancias con los resultados generales. Por ejemplo, en la oración 9 ("Ese gruñón perro prometió a esos cuñados"), con un 40 % de interpolación de inglés, el 60 % de los participantes considera que la voz pertenece a un hablante de habla anglosajona, siendo que en los resultados generales, para ese mismo grado de interpolación, menos del 25 % de los participantes lo afirmaban.

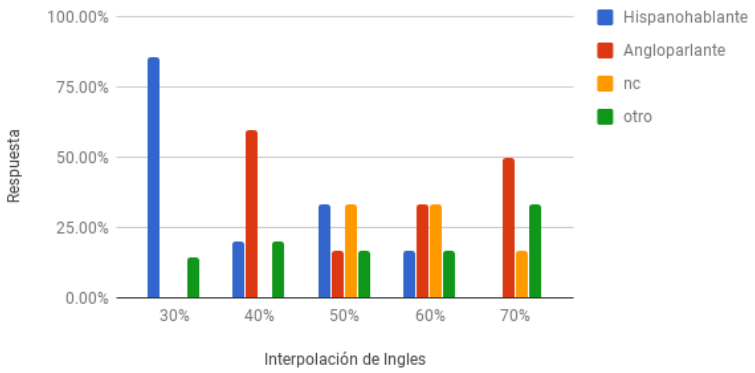


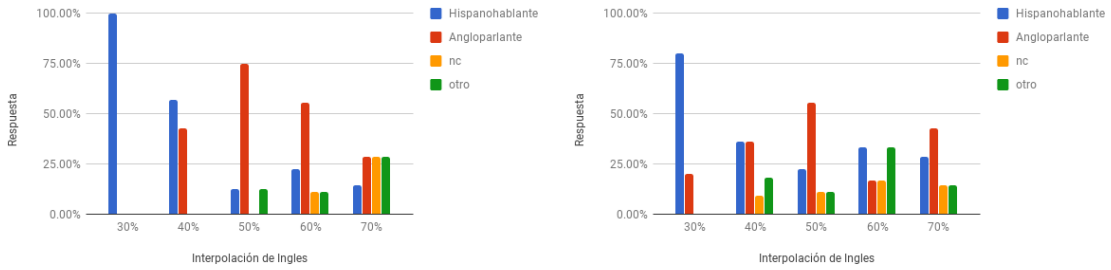

*Fig. 3.13:* Interpolación de inglés vs Respuestas de los participantes, oración 9

Buscando una explicación para este fenómeno, postulamos que esta gran diferencia entre los resultados se podría atribuir a las características particulares de cada oración. En particular la oración 9 contiene una [r] (*perro*) que resulta muy notoria al pronunciarse con una intensidad menor a la esperada (más similar a una [ɾ] (*pero*)) y puede ser atribuida, entre otras causas, a un hablante Anglosajón con dificultades en la dicción de fonos extranjeros.

Bajo esta suposición buscamos en qué otras oraciones se presenta este fono:

- Oración 1: “Mi montaña aguileña recorrió la esquina”
- Oración 6: “Su profundo riñón apoyó a Julio”
- Oración 7: “El frío churrasco oyó lo de Polonia”
- Oración 8: “Las acongojadas cotorras sonrieron a mi círculo”

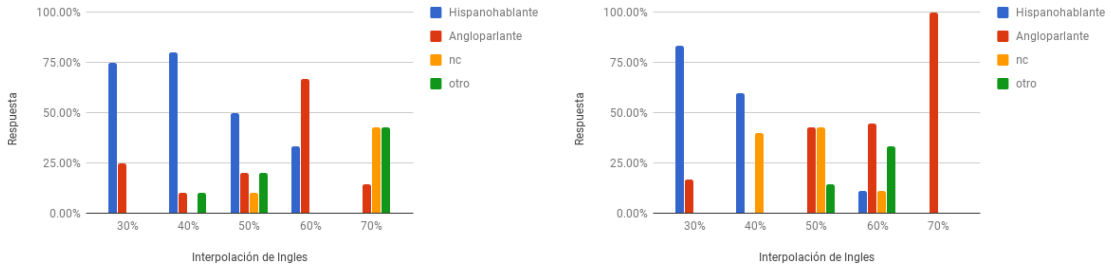
En la Figura 3.14 podemos observar que las oraciones 1 y 7 también tienen una marcada diferenciación respecto a los resultados generales. En particular estas dos oraciones no describen el comportamiento monótono observado previamente.




*(a)* Oración 1 *(b)* Oración 7

*Fig. 3.14:* Interpolación de inglés vs respuesta de los participantes

Por otro lado en la Figura 3.15 mostramos los resultados obtenidos para las oraciones 6 y 8. En estos casos no pueden observarse diferencias notorias al compararlos contra los resultados generales. Ambas oraciones presentan el mismo comportamiento monótono, que al tener una menor cantidad de datos, se observa de manera más difusa.




*(a)* Oración 6 *(b)* Oración 8

*Fig. 3.15:* Interpolación de inglés vs respuesta de los participantes

Hasta ahora analizamos los dos ejes de nuestra hipótesis por separado (por un lado, inteligibilidad, por otro, nacionalidad atribuida a la voz). En la última sección de la investigación buscaremos sacar conclusiones al componer ambos ejes en un mismo análisis.

## 3.8. Resultados generales de la experimentación

En la Figura 3.16 presentamos los resultados comparando las distancias de Levenshtein con los porcentajes de participantes que determinaron que el hablante fuera anglosajón. Al igual que en el Apartado 3.4, en el eje $x$ presentamos los grados de interpolación de inglés, yendo desde 30 % hasta 70 %. El eje $y$ del lado izquierdo representa la distancia de Levenshtein entre la oración objetivo y aquella transcripta por cada participante. Cada uno de los boxplots describe la distancia mínima obtenida y la distancia máxima (los bigotes), como así también el primer y tercer cuartil (el piso y el techo de la caja) y la mediana (línea interior que atraviesa la caja). Adicionalmente podemos observar con círculos vacíos los outliers de la muestra. En el eje $y$ del lado derecho, presentamos el porcentaje de participantes que atribuyeron la voz a un hablante anglosajón indicado mediante puntos verdes en el gráfico.

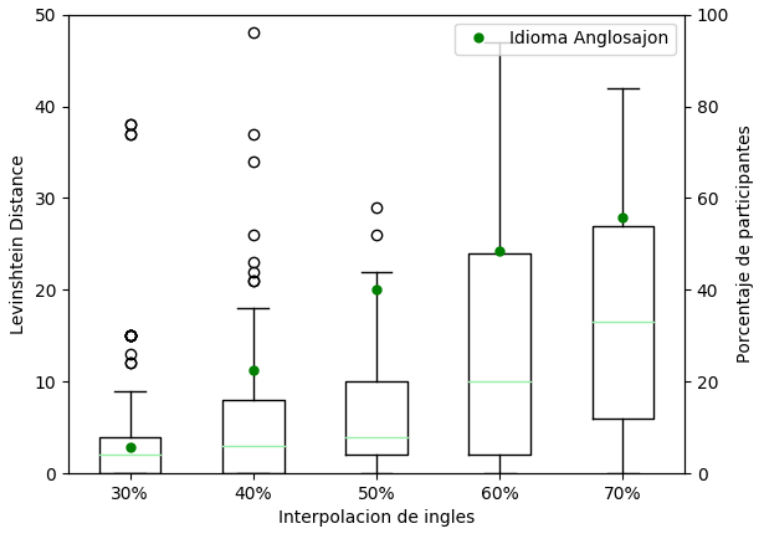



*Fig. 3.16:* Distancia de Levenshtein vs distintos grados de interpolación vs Porcentaje de participantes que reportaron hablante anglosajón

En esta figura podemos apreciar que para cada aumento en el grado de interpolación de inglés, tanto el porcentaje de participantes que considera la voz como un hablante anglosajón como la mediana de la distancia de Levenshtein crecen de manera monótona. Comenzando con un 5 % de participantes afirmando que la voz anglosajona y una mediana de la distancia de Levenshtein de 3 para una interpolación de 30 % inglés, y llegando hasta 55 % de participantes afirmando que la voz pertenecía a un hablante anglosajón y una mediana de aproximadamente 18 caracteres para la interpolación de 70 % inglés.

En base a estos resultados podemos afirmar que la hipótesis original tiene cierto grado de validez experimental: la interpolación de HMMs descripta en esta tesis es un método válido para generar voces que pueden ser identificadas con un nativo anglosajón hablando

castellano. Sin embargo esto viene con un costo asociado: la claridad de las oraciones disminuirá a medida que se aumenta el grado de interpolación del modelo de inglés.

# 4. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo tuvimos como objetivo generar un sistema de síntesis de habla en castellano con acento extranjero inglés. Para esto entrenamos de síntesis de habla con corpus en castellano e inglés, realizamos un mapeo fonético entre ambos para que fueran compatibles y luego interpolamos entre ellos para obtener distintas pronunciaciones.

Una vez que consideramos que la calidad del sistema era lo suficientemente buena, desarrollamos una encuesta online que nos permitiera probar de manera experimental que el sistema realmente cumplía con el objetivo del trabajo. Se presentaba a los participantes audios sintetizados con distintos grados de interpolación de inglés y castellano y debían determinar la nacionalidad del hablante y dar una transcripción de lo escuchado.

A través de análisis estadísticos y perceptuales de los datos pudimos concluir que existe evidencia significativa entre el porcentaje de inglés utilizado en la interpolación y la cantidad de participantes que determinan que el origen del hablante es Anglosajón.

Además, pudimos concluir que existe evidencia estadística significativa entre el porcentaje de inglés utilizado en la interpolación y el empeoramiento de la calidad de las transcripciones de los participantes.

Como fue discutido en la Sección 3.6.2 queda pendiente solucionar los problemas encontrados en el mapeo fonético. Se esperaría que un mapeo sin estos errores genere mejores resultados.

En esa misma sección también se discutió brevemente el hecho de que no todas las oraciones presentan la misma dificultad para ser comprendidas, y que esto podía deberse a que no todos los fonos pueden interpolarse de la misma manera (algunos se rompen más rápido que otros, por decirlo de alguna manera). Una posible extensión a este trabajo sería generar una interpolación controlada que permita regular cada fono por separado. Para fonos que puedan resultar problemáticos como el caso de la [r] vibrante el grado de interpolación podría dejarse más cercano al castellano, mientras que para fonos con comportamientos más similares el grado de interpolación podría llevarse más cerca del modelo inglés.

Por último, queda como trabajo futuro probar si la técnica descripta a lo largo de este trabajo es transferible a otros idiomas. Un interrogante válido es si es posible construir el sistema inverso, es decir, un sistema que sintetice habla en inglés con acento castellano/latino. Para esto, podría utilizarse una implementación similar a la propuesta en este trabajo: generar un mapeo de fonos donde todo fono del inglés esté cubierto por alguno del castellano, utilizar esto para entrenar un HMM en castellano que utilice los fonos del inglés y luego interpolar entre con este modelo y uno en inglés para obtener distintas mezclas.

# 5. APÉNDICE

## 5.1. Transcripciones ingresadas por los participantes

Oraciones originales:

| #Oración | Transcripción |
|---|---|
| 1 | Mi montaña aguileña recorrió la esquina |
| 2 | Aquel fuerte vidrio prefirió aquel botón |
| 3 | Este enjoyado juez comprará nuestro corchete |
| 4 | Tu estrecho posavasos gritó la fechoría |
| 5 | Nuestro nublado tigre concluyó a este chupetín |
| 6 | Su profundo riñón apoyó a Julio |
| 7 | El frío churrasco oyó lo de Polonia |
| 8 | Las acongojadas cotorras sonrieron a mi círculo |
| 9 | Ese gruñón perro prometió a esos cuñados |
| 10 | El nudillo argentino perdió su vaso |

Transcripciones: Oración, Mezcla de inglés, Origen, Transcripción

| | | | |
|---|---|---|---|
| 1 | 30 | Argentina | Mi montaña aguileña recorrió la esquina |
| 1 | 30 | Argentino | Mi montaña aguileña recorrio la esquina |
| 1 | 30 | española | mi montaña aguileña recorrió la esquina |
| 1 | 30 | Norte argentino | Mi montaña aguileña recorrió la esquina |
| 1 | 30 | argentina | mi montania aguilenia recorrió la esquina |
| 1 | 30 | Latino | Mi montaña aguileña recorrió la esquina |
| 1 | 40 | Español | Mi montaña aguileña recorrió la esquina |
| 1 | 40 | Ingles | MI montaña aguileña recorrió la esquina |
| 1 | 40 | Uruguay | Mi montaña recorrió la esquina |
| 1 | 40 | Argentina | mi montaña aguileña recorrió la esquina |
| 1 | 40 | centroamerica | Mi montaña aguileña recorrió la esquina |
| 1 | 40 | estadounidense | mi montaña aguileña recorrio la esquina |
| 1 | 40 | EEUU | Mi montaña aguileña recorrió la esquina |
| 1 | 50 | Boliviana | Mi montaña recorrio la esquina |
| 1 | 50 | EEUU hablando español | mi montaña milenia recorrió la esquina |
| 1 | 50 | Inglés | Mi montaña aguileña recorrio la esquina |
| 1 | 50 | Japonés | Recorrio la esquina |
| 1 | 50 | norteamericana | mi montaña aguileña recorrió la esquina |
| 1 | 50 | Inglesa | mi montaña pedigüeña recorrió la esquina |
| 1 | 50 | Estadounidense | mi montaña aguileña recorrio la esquina |
| 1 | 50 | Estadounidense | mi montaña aguileña recorrió la esquina |
| 1 | 60 | estadounidense | Mi montaña aguileña recorrió la esquina |
| 1 | 60 | Estados unidos | Mi montaña aguilenia recorrió la esquina |
| 1 | 60 | Irlanda | Mi montaña aguileña recorrió la esquina. |
| 1 | 60 | Norte argentino | Mi montaña aguileña recorrió la esquina |
| 1 | 60 | catellano | recorriendo la esquiena |
| 1 | 60 | EEUU | mi montaña aguileña recorrió la esquina |
| 1 | 60 | No se especifica | Mi montaña..... recorrió la esquina |
| 1 | 60 | eeuu | mi montaña aguileña recorrió la esquina |
| 1 | 60 | frances | mi montaña recorrió la esquina |
| 1 | 70 | Ninguna | Mi monzonia la cigüeña recorrio la esquina |
| 1 | 70 | Italiano | Corrio a la esquina |
| 1 | 70 | Castellano | Recorrió la esquina |
| 1 | 70 | Guam | Mí montoña aguileña recorrió la esquina |
| 1 | 70 | Canadá | Mi montaña aguileña recorrió la esquina |
| 1 | 70 | EEUU | Mi corrió la esquina |
| 1 | 70 | nose | |
| 2 | 30 | argentino | Aquel fuerte vidrio prefirió aquel botón |
| 2 | 30 | Ninguna | Aquel fuerte vidrio prefirio aquel boton |
| 2 | 30 | Ninguna | Aquel fuerte vidrio prefirió aquel botón |
| 2 | 30 | Argentino | aquel fuerte vidrio prefirio aquel boton |
| 2 | 30 | Chile | aquel fuerte vidrio prefirió aquel botón |
| 2 | 30 | Argentina | aquel fuerte vidrio prefirió aquel botón |
| 2 | 40 | Espa;ol | Aquel fuerte vidrio prefirio aquel boton |
| 2 | 40 | Ingles | Aquie fuerte vidrio prefirio aquel boton |
| 2 | 40 | Argentino | Aquel fuerte vidrio prefirio aquel boton |
| 2 | 40 | Norte argentino | Aquel fuerte vidrio prefirió aquel botón |

| | | | |
|---|---|---|---|
| 2 | 40 | centroamericana | aquel fuerte vidrio prefirió aquel boton |
| 2 | 40 | argentina | aquel fuerte vidrio percibió al boton |
| 2 | 40 | Gallega | Aquel fuerte vidrio prefirió aquel botón. |
| 2 | 50 | Inglesa | Aqueo fuerte vidrio perforo aquel boton |
| 2 | 50 | paraguay | aquel vidrio presiono aquel boton |
| 2 | 50 | francés | aquél fuerte vidrio prefirió aquél botón |
| 2 | 50 | Estados Unidos | aquel fuerte vidrio percibió aquel botón |
| 2 | 50 | EEUU | Aquel fuerte vidrio prefirió aquel botón |
| 2 | 50 | Mexicano | Aquel fuerte vidrio percibió aquel botón |
| 2 | 60 | inglesa | aquel fuerte vidrio presidió aquel botón |
| 2 | 60 | Santo Domingo | Aquel Fuel de vidrio prefiere aquel boton |
| 2 | 60 | paraguay | te voy a pegar un palazo en la nuca y romperte el orto |
| 2 | 60 | estadounidense | prefirió el botón |
| 2 | 60 | China | |
| 2 | 70 | Ingles | |
| 2 | 70 | Estados unidos | Preferido aquel boton |
| 2 | 70 | inglesa | aquel blablaba prefirio aquel boton |
| 2 | 70 | No se especifica | Aquel fuerte vidrio prefirió aquel botón |
| 2 | 70 | Ninguna, es un robot. | |
| 2 | 70 | EEUU | Aquel fuerte vidrio prefirió aquel boton |
| 3 | 30 | Argentino | Este enjollado puede comparar nuestro cochete |
| 3 | 30 | Argentina | Este enfueyado fue comparado corchete |
| 3 | 30 | Español | Corchete |
| 3 | 30 | Argentino capital federal | Corchete |
| 3 | 30 | Robot | Este enrollado fue comprado nuestro corchete |
| 3 | 30 | española | este enjoyado juez comprará nuestro corchete |
| 3 | 30 | Paraguayo | este enjoyado fue comprado este corchete |
| 3 | 30 | Argentino | Este enfollado juez comprará nuestro corchete |
| 3 | 30 | Español | Este enfollado fue comprar nuestro corchete |
| 3 | 30 | No se especifica | Este enjoyado juez comprará nuestro corchete |
| 3 | 30 | Indefinida | Este enjoyado fue comprado corchete |
| 3 | 40 | La Rioja, Argentina | Este encollado encontrará nuestro corchete |
| 3 | 40 | Ninguna nacionalidad | Este fue juez corchete |
| 3 | 40 | España | este enfollado comprara nuestro corchete |
| 3 | 40 | Española | este fue nuestro corchete |
| 3 | 40 | Español neutro | este enjoyado juez comprará nuestro corchete |
| 3 | 50 | Boliviano | Este juez comprara nuestro |
| 3 | 50 | anglo | este enrollado corchete |
| 3 | 50 | Maracaibo | Este enconchado puede comprado nuestro cohete |
| 3 | 50 | Colombiano/Venezolano | Este enfollado fue encontrar nuestro corchete |
| 3 | 50 | No se | Este enrollado fue encontrar nuestro coechete |
| 3 | 60 | Ninguna | Este en...fue comprar nuestro corchete |
| 3 | 60 | Inglesa | Este nuestro corchete |
| 3 | 60 | gallego | este enfollado comprara nuestro corchete |
| 3 | 60 | Ingles | Este enfollado juez comprara nuestro corchete |
| 3 | 60 | inglés | compraran nuestro corchete |
| 3 | 60 | Español | Este ??? pues comprará nuestro corchete |
| 3 | 60 | estadounidense o inglés | este ... nuestro corchete |

| | | | |
|---|---|---|---|
| 3 | 60 | estadounidense | este fue nuestro corchete |
| 3 | 60 | Argentino | Esté nuestro corchete |
| 3 | 60 | americano (EEUU) | estoy nuestro corchete |
| 3 | 60 | Desconocido | Comprara nuestro corchete |
| 3 | 70 | alemania | este enfollado fuescon cuero nuestro follete |
| 3 | 70 | Hablante nativo de inglés | este ..... fue ..... en nuestro corchete |
| 3 | 70 | estadounidense | este vuestro corchete |
| 3 | 70 | Ruso | este fue encontrado con nuestro coche |
| 3 | 70 | estadounidense | nuestro corchete |
| 4 | 30 | Argentina | Tu estrecho posa basos grito la fechoria |
| 4 | 30 | Argentina | Tu estrecho posa vasos grito |
| 4 | 30 | Cigbord | Gritemos portavasos gritona fechoria |
| 4 | 30 | parece | una voz artificial Tu estrecho posavasos gritó la fechoría |
| 4 | 30 | españa | tu estrecho posavasos gritó la fechoría |
| 4 | 30 | Española | Tu estrecho posavasos gritó la fechoría |
| 4 | 30 | no sé | su estrecho posavasos gritó en la fechoría |
| 4 | 30 | Española | Tu estrecho posa vasos gritó la fechoría |
| 4 | 30 | centroamericano | fui estrecho posavasos grito la fechoría |
| 4 | 30 | español neutro | tu estrecho posavasos gritó la fechoría |
| 4 | 40 | español de españa | tu estrecho posavasos grito la fechoria |
| 4 | 40 | España | Tu estrecho posavasos gritó la fechoría |
| 4 | 40 | Español | Si estrecho posa vaso |
| 4 | 40 | España | tu estrecho posavasos gritó la fechoría |
| 4 | 40 | Argentina | Tu estrecho posa vasos grito la fechoriay |
| 4 | 50 | anglo | tu estrecho posa vasos grito la fechoria |
| 4 | 50 | Ingles | Su estrecho posavazos gritó la fechoría |
| 4 | 50 | Ingles | Su estrecho posavazos gritó la fechoría |
| 4 | 50 | italiano | la fechoria |
| 4 | 50 | inglés | su estrecho posavasos gritó la fechoria |
| 4 | 50 | estadounidense | tu estrecho posavasos gritó la fechoría |
| 4 | 50 | Cuba | tu estrecho posavasos, grito la fechoría |
| 4 | 50 | Española | Su estrecho posavaso fechoría |
| 4 | 60 | Francés | Frechoia |
| 4 | 60 | chino mandarin | chu ... |
| 4 | 60 | Colombia | Tu estrecho posavasos grito la fechoria |
| 4 | 60 | Eeuu | Posavasos |
| 4 | 60 | estaunidense | ... la fechoría |
| 4 | 70 | estadounidense | la fechoria |
| 4 | 70 | Estados | Unidos tu estrecho posavasos |
| 4 | 70 | no lo pude determinar, osea es una maquina | las fechorias (al final de todo) |
| 4 | 70 | Australia | You |
| 4 | 70 | ns/nc | ns/nc |
| 4 | 70 | inglesa | su estrecho posavasos gritó la fechoría |
| 4 | 70 | Hablante nativo de inglés | tu estrecho posavasos .. la fechoría |
| 4 | 70 | americana (EEUU) | tu estrecho portavasos fechoria |
| 4 | 70 | estadounidense | vasos la fechoría |
| 5 | 30 | Castellano | Nuestro nublado tigre concluyo a este chupetin |
| 5 | 30 | Española | Nuestro nublado tigre concluyó a esta chupetin |
| 5 | 30 | Sin nacionalidad | Nuestro nublado tigre concluyó a este chupetin |
| 5 | 30 | argentina | nuestro nublado tigre concluyó a este chupetín |
| 5 | 30 | Neutro | Nuestro nublado tigre concluyó a este chupetin |

| | | | |
|---|---|---|---|
| 5 | 30 | argentina | nuestro nublado tigre concluyó a nuestro chupetín |
| 5 | 40 | Argentina | Nuestro nublado día concluyó a este chupetin |
| 5 | 40 | Alemana | Nuestro nublado vive concluyó este chupetin |
| 5 | 40 | Inglesa | nuestro nublado y reconcluyó a este chupetín |
| 5 | 40 | EEUU | nuestro nublado tigre concluyo este chupetin |
| 5 | 40 | Argentino | nuestro nublado tigre concluyó a este chupetín |
| 5 | 40 | español | Nuestro nublado tigre concluyó a este chupetín |
| 5 | 50 | ciudad de buenos aires | nuestro nublado concluyo este chupetin |
| 5 | 50 | Google | Nuestro nublado vive concluyo a este chupetín |
| 5 | 50 | estadounidense | nuestro nublado tigre concluyó este chupetín |
| 5 | 50 | Robot del traductor de Google | Nuestro hermano encontró este chupetín (?) |
| 5 | 50 | inglés | Nuestro nublado día concluyó este chupetín |
| 5 | 50 | estaunidense | nuestro nublado tigre concluyó este chupetín |
| 5 | 50 | Robot | . Nuestro nublado concluyó este chupetín |
| 5 | 50 | español | neutro nuestro nublado ... concluyó a este chupetín |
| 5 | 50 | Argentina | Nuestro nublado y reconstruyó a este chupetín |
| 5 | 60 | Español | |
| 5 | 60 | Chile | nuestro ... chupetin |
| 5 | 60 | español | este chupetin |
| 5 | 60 | Estados Unidos | nuestro nublado concluyó este chupetín |
| 5 | 60 | Eeuu | Nuestro nublado y re concluyó este chupetin |
| 5 | 60 | Español | neutro Nuestro ... este chupetin |
| 5 | 60 | Inglés/Estado Unidense | Nuestro nublado tigre concluyó este chupetin |
| 5 | 60 | estadounidense o inglés | nuestro nublado y reconcluyó este chupetín |
| 5 | 70 | polaco | nuestro reconstruye este chupetìn |
| 5 | 70 | Portugués | Nuestro pequin |
| 5 | 70 | Estadounidense | Nuestro nublado construyó este chupetin |
| 5 | 70 | Estados unidos | Nuestro nublado libre construyó este chupetín |
| 5 | 70 | Estadounidense | nuestro ... este chupetín |
| 5 | 70 | Panama | nuestro anublado reconcluyo este chupetin |
| 5 | 70 | Argentina | Nuestro nublado concluyó a este chupetín |
| 5 | 70 | estadounidense | nuestro nublado concluyó este chupetín |
| 6 | 30 | Español | Su profundo riñón apoyo a Julio |
| 6 | 30 | Argentino | Su profundo riñón apoyó a Julio |
| 6 | 30 | Argentina | su profundo riñon apoyo a Julio |
| 6 | 30 | estados unidos | su profundo riñón apoyó a Julio |
| 6 | 30 | argentina | su profundo riñón apoyó a julio |
| 6 | 30 | Española | Su profundo riñón apoyó a a Julio |
| 6 | 30 | Argentino | Su profundo riñón apoyó a Julio |
| 6 | 30 | Ingles | Su profundo riñón apoyo a Julio |
| 6 | 40 | Colombiano | Su profundo riñón apoyo a julio |
| 6 | 40 | frances | nada |
| 6 | 40 | argentina | si lo profundo julio |
| 6 | 40 | Paraguayo | Profundo riñon julio |
| 6 | 40 | Argentino | Su profundo riñón apoyo a julio |
| 6 | 40 | Español neutro/-Rioplatense | Su profundo riñon apoyo a Julio |
| 6 | 40 | EEUU | su profundo riñon apoyo junio |
| 6 | 40 | Argentina | su profundo riñon apoyó a julio |
| 6 | 40 | Argentino | Su profundo riñón apoyo a julio |
| 6 | 40 | Argentina | Su profundo riñón apoyó a Julio |
| 6 | 50 | polaco | si profundo donde apoyo a julio |
| 6 | 50 | Castellano | Su profundo riñon apoyo a julio |
| 6 | 50 | Argentino | Su profundo riñon julio |
| 6 | 50 | Francia | su profundo riñor apoyo a |
| 6 | 50 | no es hispano parlante | su profundo riñon de apoyo julio |

| | | | |
|---|---|---|---|
| 6 | 50 | Peru | su profundo riñon apoyo a Julio |
| 6 | 50 | español centroamérica | su profundo riñon apoyó a Julio |
| 6 | 50 | Inglesa | su profundo riñón apoyó a Julio |
| 6 | 50 | Argentina | Su profundo riñón apoyo a julio |
| 6 | 50 | estadounidense | Su profundo riñón apoyó a julio |
| 6 | 60 | estados unidos | su profundo riñon apoyo a julio |
| 6 | 60 | Estadounidense | su profundo riñón apoyó a Julio |
| 6 | 60 | Estadounidense | Su profundo riñón apoya a julio |
| 6 | 60 | Argentina | Su profundo riñón apoyó a Julio |
| 6 | 60 | cordobés | su profundo riñón apoyó a julio |
| 6 | 60 | estadounidense | su profundo riñón apoyó a julio |
| 6 | 70 | Google | Su profundo riñón apoyó a su higado |
| 6 | 70 | Rusia | |
| 6 | 70 | google translate | supra india fidileia |
| 6 | 70 | estadounidense | su profundo riñón apoyo afilio |
| 6 | 70 | Robot | Su profundo |
| 6 | 70 | Japones | Si profundo neanea nou |
| 6 | 70 | china | su profundo riñon apoyó a julio |
| 7 | 30 | Estadounidense | el frío churrasco yo lo de Polonia |
| 7 | 30 | Argentina | El frío churrasco oyó lo de Polonia |
| 7 | 30 | Colombia | el frio churrasco de colombia |
| 7 | 30 | argentina | el frío churrasco lleno de polonia |
| 7 | 30 | chile | el frio solo de polonia |
| 7 | 40 | Español | Churrasco, Polonia |
| 7 | 40 | estados unidos | el frio churrasco de polonia |
| 7 | 40 | español | |
| 7 | 40 | Estadounidense | Enfrío churrasco "goyono"de Polonia |
| 7 | 40 | Española | el frio churrasco frio de Polonia |
| 7 | 40 | Español | Enfrió churrasco en solo de polonia |
| 7 | 40 | ni idea'); | el frio churrasco oyó lo de Polonia |
| 7 | 40 | Inglesa | EL frío churrasco ozono de Polonia |
| 7 | 40 | estadounidense | el frío churrasco oyó lo de Polonia |
| 7 | 40 | Neerlandes | Enfrio churrasco yo no de polonia |
| 7 | 40 | Japones | Enfrió churrasco frío de polonia |
| 7 | 50 | Español | Enfrio churrascos anchos de Polonia |
| 7 | 50 | Inglesa | El frío churrasco yo como de colombia |
| 7 | 50 | Inglés o Irlandés | churrasco Polonia |
| 7 | 50 | computadora | en frio churrasco yo no de polonia |
| 7 | 50 | EEUU | Enfrío churrasco lleno de poloña. |
| 7 | 50 | Estados unidos | enfrío churrasco oyó lo de polonia |
| 7 | 50 | Brasiltino | Enfrio churrasco o jogo de Polonia |
| 7 | 50 | Argentina | el frio churrasco ollolo de polonia |
| 7 | 50 | Ingles | Enfrió churrasco de polonia |
| 7 | 60 | Castellano | Frio churrasco lleno de colonia |
| 7 | 60 | Argentino | El frío churrasco de Polonia |
| 7 | 60 | Polonia | El frío churrasco |
| 7 | 60 | Francia | el frio churrasco de polonia |
| 7 | 60 | Estados Unidos (google translate...) | el frio churrasco llego de polonia |
| 7 | 60 | ninguna | el frío churrasco moncholo de polonia |
| 7 | 70 | Google | El frio churrasco de polonia |
| 7 | 70 | inglés | polonia |
| 7 | 70 | eeuu | el frio churrasco de polonia |
| 7 | 70 | EEUU | El frio churrasco de polonia |
| 7 | 70 | Argentina | El frío churrasco cayo de Polonia |
| 7 | 70 | Quichua | Churrasco |
| 7 | 70 | Argwntino | El frió churrasco lleno de Colonia |
| 8 | 30 | España (Sur) | Las acongojadas cotorras sonrieron a mi círculo. |
| 8 | 30 | EEUU | Las acongojadas cotorras sonrieron a mi círculo. |
| 8 | 30 | argentina | Lss acongojadas cotorras sonrieron a circulo |
| 8 | 30 | Argentina | Las acongojadas cotorras sonrieron a mi círculo |
| 8 | 30 | Latino | Las acongojadas cotorras sonrieron a mi circulo |
| 8 | 30 | Español | Las acongojadas cotorras corrieron a mi circulo |
| 8 | 40 | Español | Las acontojadas culturas Sonrieron en semicirculo |

| | | | |
|---|---|---|---|
| 8 | 40 | Argentino | Las acombojadas cotorras sonrieron a mi círculo |
| 8 | 40 | Google | Las acongojadas cotorras sonrieron a mi circulo |
| 8 | 40 | Latino | Las acongojadas cotorras sonrieron a mi círculo |
| 8 | 40 | asd | asd |
| 8 | 50 | Francia | las cotorras sonrieron en mi circulo |
| 8 | 50 | Nose | Las acongojadas cotorras sonrieron a mi circulo |
| 8 | 50 | Ninguno | Las acongojadas cotorras sonrieron a mi círculo |
| 8 | 50 | Estadounidense | Las acongojadas cotorras vinieron a mi círculo |
| 8 | 50 | Ingles/estadounidense | Las acongojadas cotorras sonrieron a mi circulo |
| 8 | 50 | ninguna | las acongojadas cotorras sonrieron a mi círculo |
| 8 | 50 | estadounidense | las aconogojadas cortorras se unieron a mi círculo |
| 8 | 60 | Español | De mi circulo |
| 8 | 60 | Estadounidense | En mi circulo |
| 8 | 60 | Estadounidense | Las acongojadas cotorras sonrieron a mi círculo |
| 8 | 60 | Inglesa | Mi círculo |
| 8 | 60 | Rusa | Las acongojadas cotorras sonrieron a mi círculo |
| 8 | 60 | EEUU | las acongojadas cotorras sonrieron en mi circulo |
| 8 | 60 | No hispanohablante | las acongoyadas cotorras sonrieron a mi circulo |
| 8 | 60 | rusa | las acongojadas cotorras sonrieron en mi circulo |
| 8 | 60 | Francesa | Las acongojadas cotorras sonrieron en mi círculo |
| 8 | 70 | estadounidense | círculo |
| 8 | 70 | estadounidense | some hemicírculo |
| 8 | 70 | inglés | sonrieron en mi circulo |
| 8 | 70 | Estadounidense | ... sonrieron en mi círculo |
| 8 | 70 | eeuu | plaza sombreada con sombrero sonrieron en mi círculo |
| 9 | 30 | Español | Ese gruñon perro prometio a esos cuñados |
| 9 | 30 | Argentina | Ese gruñón Prometió a sus cuñados |
| 9 | 30 | Paraguayo | Ese gruñon me lo prometio a esos cuñados |
| 9 | 30 | español | ese gruñon se lo prometio a esos cuñados |
| 9 | 30 | Argentino | el se reunión pero prometió a esos cuñados |
| 9 | 30 | Rusa | Ese gruñón perro prometió a esos cuñados |
| 9 | 30 | Argentina | Ese gruñon perro prometio esos cuñados |
| 9 | 40 | Francia o Suiza | Ese gruñón perro prometió a esos cuñados. |
| 9 | 40 | estadounidense | Ese gruñón perro prometió a esos cuñados |
| 9 | 40 | Chilena | Ese gruñón perro prometió a esos cuñados |
| 9 | 40 | estadounidense | ese gruñón perro prometió a esos cuñados |
| 9 | 40 | eeuu | ese grunion perro prometió a esos cuñados |
| 9 | 50 | Google | Ese gruñon perro prometio a esos cuñados |
| 9 | 50 | puerto rico | ese gruñón perro prometió esos cuñados |
| 9 | 50 | Estados unidos | pero prometió a esos cuñados |
| 9 | 50 | es un robot | el segundo *** prometio a esos cuñados |
| 9 | 50 | ucrania | ese gruñon perro prometió a esos cuñados |
| 9 | 50 | Mexico | ese gruñón perro prometió a esos cuñados |
| 9 | 60 | Sueco | Ese gruñon prometio esos cuñados |
| 9 | 60 | españa | Ese gruñón peón prometió a esos cuñados |
| 9 | 60 | a | esas cuñadas |
| 9 | 60 | H | J |
| 9 | 60 | Ingles | Esa reunion prometiò a esos cuñados |

| | | | |
|---|---|---|---|
| 9 | 60 | Anglo | parlante Ese gruñón perro a esos cuñados |
| 9 | 70 | Ingles | Its a pair of promises |
| 9 | 70 | Estadounidense | Ese gruñón pero prometió a esos cuñados |
| 9 | 70 | Estadounidense | Ese gruñón pero prometió a esos cuñados |
| 9 | 70 | ... | ese gruñón perro prometio a esos cuñados |
| 9 | 70 | Aleman | ese gruñon perro prometió a esas cuñadas |
| 9 | 70 | Ruso | Rsts reunion que prometió a sus... |
| 10 | 30 | Español | El nudillo argentino perdio su vaso |
| 10 | 30 | Argentino | El nudillo argentino perdio su vaso |
| 10 | 30 | Argentina | El novillo argentino perdió su vaso |
| 10 | 30 | Español | El nudillo argentino perdio su vaso |
| 10 | 30 | argentina | el nudillo argentino perdió su vaso |
| 10 | 30 | Argentina | el nudillo argentino perdió su vaso |
| 10 | 40 | - | El argentino perdió su vaso |
| 10 | 40 | ... | el nudillo argentino perdió su vaso |
| 10 | 40 | argentina | el nudillo argentino perdio su vaso |
| 10 | 40 | argentino | el argentino perdió su vaso |
| 10 | 40 | ingles | el nudillo argentino perdió su vaso |
| 10 | 40 | Argentino | El....perdió su vaso |
| 10 | 50 | Español | El nudillo argentino perdió su vaso |
| 10 | 50 | Mexico | el novillo argentino perdió su vaso |
| 10 | 50 | argentina | el novillo argentino perdió su vaso |
| 10 | 50 | español | el brillo argentino perdio su vaso |
| 10 | 50 | No se especifica | Individuo argentino perdió su vaso |
| 10 | 50 | Ingles | El novillo argentino perdió su vaso |
| 10 | 50 | .? | El argentino perdió su vaso |
| 10 | 60 | Estados unidos | El Argentino perdio su vaso |
| 10 | 60 | estadounidense | El ... argentino perdió su vaso |
| 10 | 60 | robot | el nudillo argentino perdio su vaso |
| 10 | 60 | EEUU | argentino perdió su vaso |
| 10 | 60 | amerciano (EEUU) | un chico argentino perdió su vaso |
| 10 | 70 | estadounidense | El chico argentino perdió su vaso |
| 10 | 70 | estadounidense | el nuevo dj argentino perdió su vaso |
| 10 | 70 | Taiwán | El ... argentino perdió su vaso. |
| 10 | 70 | Estadounidense | El nuevo argentino perdio su vaso |
| 10 | 70 | Computadorlandia | El indígena argentino perdió su vaso |
| 10 | 70 | estadounidense | argentino perdió su vaso |
| 10 | 70 | estadounidense | el susuño argentino perdió su vaso |
| 10 | 70 | Estados Unidos | ... argentino perdio su casa |

## Bibliografía


[1] S. O. Arik & M. Chrzanowski & A. Coates & G. Diamos & A. Gibiansky & Y. Kang & X. Li & J. Miller & J. Raiman & S. Sengupta & M. Shoeybi. Deep Voice: Real-time neural text-to-speech. International Conference on Machine Learning, 2017.

[2] Sotelo, Jose & Mehri, Soroush & Kumar, Kundan & Santos, Joao Felipe & Kastner, Kyle & Courville, Aaron & Bengio, Yoshua, Çhar2wav: End-to-end speech synthesis". International Conference on Learning Representations, 2017.

[3] Yamagishi, Junichi & Onishi, Koji & Masuko, Takashi & Kobayashi, Takao. (2005). Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. IEICE Transactions on Information and Systems. E88D. 10.1093/ietisy/e88-d.3.502.

[4] Nose, Takashi & Yamagishi, Junichi & Masuko, Takashi & Kobayashi, Takao. (2007). A Style Control Technique for HMM-Based Expressive Speech Synthesis. IEICE Transactions. 90-D. 1406-1413. 10.1093/ietisy/e90-d.9.1406.

[5] Boersma, Paul (2001). Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

[6] http://hts.sp.nitech.ac.jp/?Download

[7] Torres, Humberto M., & Jorge A. Gurlekian. ."Automatic determination of phrase breaks for Argentine Spanish."Speech Prosody 2004, International Conference. 2004.

[8] Torres, Humberto M. & Jorge A. Gurlekian & C. C. Mercado. ."Aromo: Argentine spanish TTS system."Proc. VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH.

[9] Kominek, John & W Black, Alan. (2004). The CMU Arctic speech databases. SSW5-2004.

[10] Zen, H. & Braunschweiler, N. & Buchholz, S. & Gales, M. J. & Knill, K. & Krstulovic, S. & Latorre, J. (2012). Statistical parametric speech synthesis based on speaker and language factorization. IEEE transactions on audio, speech, and language processing, 20(6), 1713-1724.

[11] Yoshimura, Takayoshi. "Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems."Nagoya Institute of Technology, Japan (2002).

[12] http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo_CMU-ARCTIC-ADAPT.tar.bz2

[13] Wester, M., & Karhila, R. (2011, May). Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 5372-5375). IEEE.

[14] Prahallad, K. & Black, A. W. & Mosur, R. (2006, May). Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on (Vol. 1, pp. I-I). IEEE.

[15] Goldstein, B. (2001). Transcription of Spanish and Spanish-influenced English. Communication Disorders Quarterly, 23(1), 54-60.

[16] Yoshimura, T. & Masuko, T. & Tokuda, K. & Kobayashi, T. & Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In Fifth European Conference on Speech Communication and Technology.

[17] http://www.cstr.ed.ac.uk/downloads/festival/2.4/

[18] http://festvox.org/download.html

[19] http://www.cstr.ed.ac.uk/projects/speech_tools