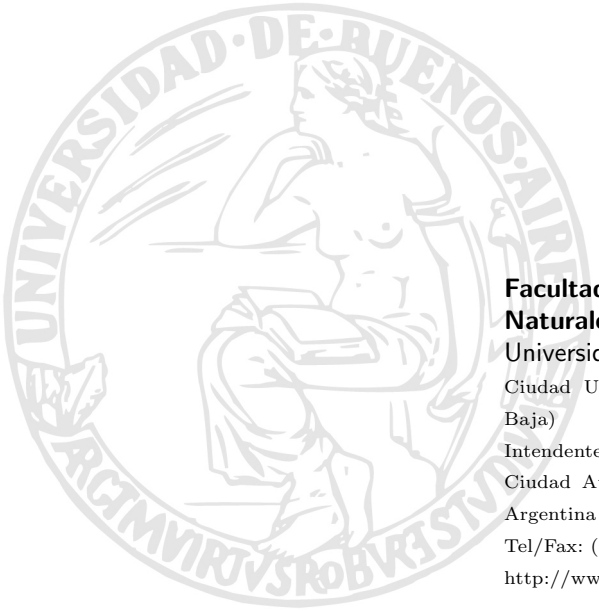
DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

On TTS with non native prosody: a systematic aproach

2 de febrero de 2018

Integrante	LU	Correo electrónico
Negri, Franco	893/13	franconegri2004@hotmail.com



Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (++54 +11) 4576-3300

http://www.exactas.uba.ar

Índice

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 muchas aplicaciones domesticas, desde navegación por GPS, asistentes personales inteligentes (como es el caso de SIRI), ayuda para personas no videntes, traducción automática, etc.

En las últimas décadas se han visto grandes progresos en este campo, siendo capaces de modelar con cierto grado de efectividad cuestiones tales como la prosodia del hablante, emociones, etc. Una técnica bastante utilizada es la que utiliza modelos ocultos de markov (HMMs) que, a partir de un corpus de datos de entrenamiento, extrae información acústica y genera un modelo probabilístico que permita sintetizar habla.

En la actualidad el campo de la síntesis utilizando HMMs presenta algunos interrogantes con respecto al entrenamiento y utilización de corpus de datos con hablantes de distintas lenguas [1,2].

En este trabajo de tesis se pretende presentar una manera posible de generar un TTS basado en HMMs capaz de sintetizar habla en español con acento extranjero. Las razones por las que podría querer diseñarse un sistema con estas características varían desde un punto de vista puramente técnico, ya que un sistema así permitiría la utilización de corpus de entrenamiento de hablantes no nativos para la generación de una nueva voz, hasta cuestiones lingüísticas, como es poder vislumbrar el limite en que un acento deja de parecernos local para pasar a ser extranjero.

En el transcurso de este trabajo se espera además evaluar la prosodia y la fonética del modelo generado con estas características, como así también evaluar su inteligibilidad. Además pretenderemos evaluar la efectividad de técnicas de speaker adaptation cuando se utilizan corpus de distintas nacionalidades con repertorios fonéticos muy disimiles (para este caso de estudio: castellano e ingles)

Para este trabajo nos basaremos fuertemente en la síntesis/análisis melcepstral, speech parameter modeling usando HMMs y speech parameter generation usando HMMs, como es descripto en la disertación doctoral Simultaneous Modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems del Professor Tadashi Kitamura del Nagoya Institute of Technology[3].

También se utilizarán las herramientas para la investigación y generación de nuevas voces Festival y Festvox, para el preprocesamiento de datos.

2. Metodología

2.1. Herramientas

En esta sección presentamos las herramientas utilizadas para este trabajo de investigación.

2.1.1. Festival y Festvox

Festival es un framework que permite sistetizar audios, como así tambien una gran variedad de APIs, para el procesamiento de audios y generación de nuevos TTS.

Festvox a su vez expande sobre Festival, agregando todavía mas herramientas relacionadas, que van desde la generación de modelos prosodicos, hasta etiquetado automatico de corpus.

Para este trabajo utilizaremos Festival y festvox para generar los Utternaces requeridos tanto para el entrenamiento como para la sintesis de audios. Estos consisten basicamente en una transcripción fonetica de los audios dividida en segmentos temporales y datos contextuales tales como la cantidad de silabas en la palabra siendo transcripta, fonemas que preceden y proceden al actual, etc.

En particular, Festival tambien cuenta con herramientas de etiquetado automatico. Para este trabajo utilizaremos EHMM alingnment, que a partir de un corpus y sus transcripciones, permite generar utternaces segmentos coinciden con aquellos de los audios.

Estos utternaces serán posteriormente utilizados en el entrenamiento con HTS para modelar cada uno de los fonemas con una mezcla de variables aleatorias gaussianas.

2.1.2. HTS

HTS es un TTS basado en HMMs que modela simultáneamente la duración, el espectro (mel-cepstrum) y la frecuencia principal ($f0$) de utilizando un framework de HMM:

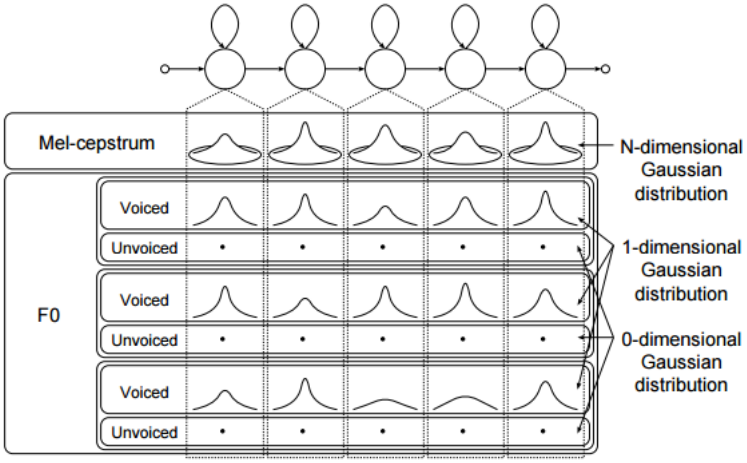


Figure 5.3: structure of HMM.

Figura 1: Estructura de un hmm (simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura, january 2002)

Por otro lado 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 extraida de los audios de entrenamiento. A continuación se presenta una vista esquemática de la estructura del HMM generado:

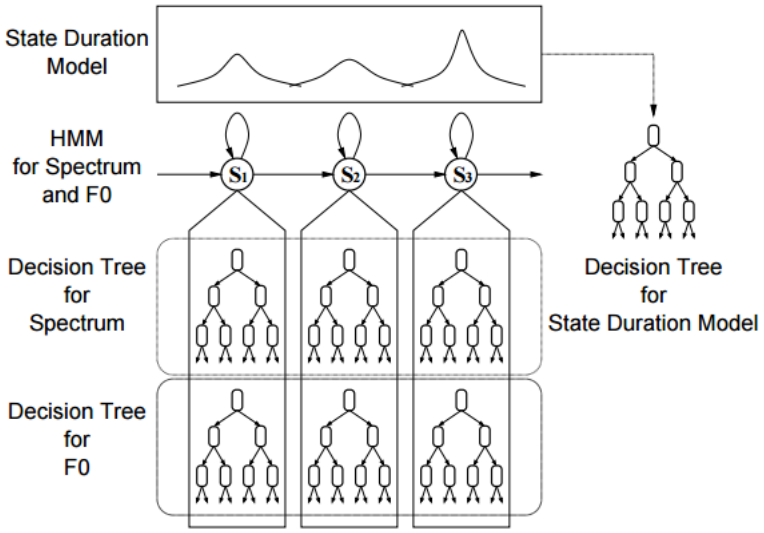


Figure 5.4: Decision trees.

Figura 2: HMM generado (simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura, january 2002)

En particular para este trabajo la clusterización de datos se realizó generando arboles de decición, para cada fonema se tomaron los dos fonemas precedentes y los dos fonemas procendetes y se extrageron las siguientes features:

- Modo de articulación del fonema.
- Punto de articulación del fonema.
- La perspectiva articulatoria (anterior, central o posterior).
- Si el fonema es una vocal o una consonante.
- En caso de ser una vocal, a que categoría pertenecía: por ejemplo para el fonema $/i/ : i, i0, i1$.
- En caso de ser una vocal, su Redondeamiento vocálico.
- En caso de ser una consonante, si es lennis o fortis.

En la siguiente imagen se muestra un fragmento de un arbol de decición generado para modelar la duración de los fonemas en un hmm:

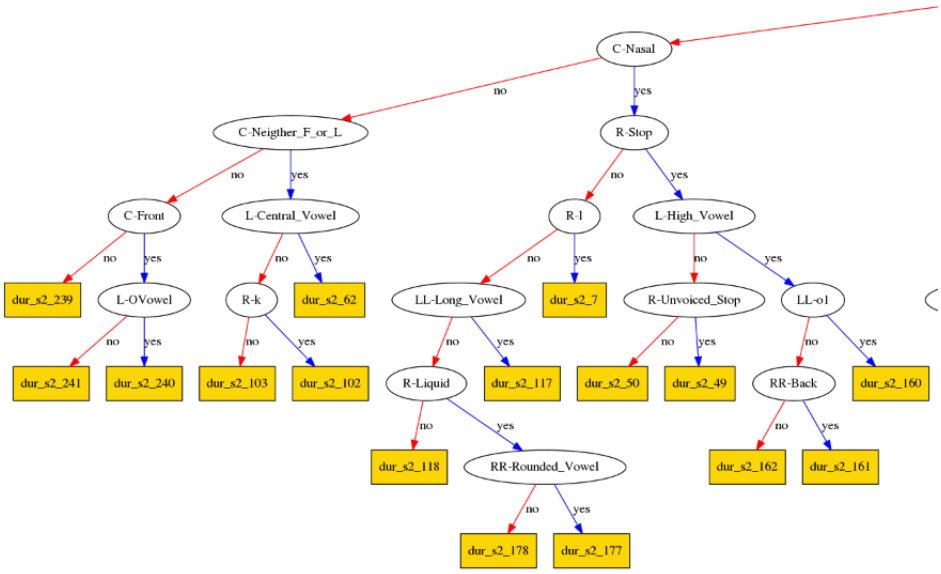


Figura 3: Árbol de decición generado

Con este modelo, el sistema podra inferir por ejemplo cosas como: si el fonema actual no es nasal (C-Nasal) seguido de un stop (R-Stop), que no es el fonema l estará modelado por función de probabilidad gaussiana definida en dur_s2_7.

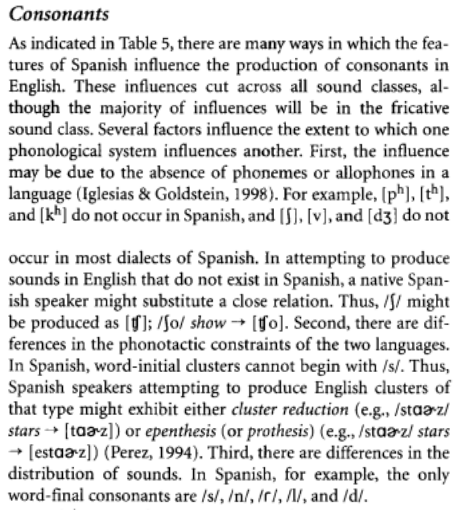
En las primeras iteraciones del desarrollo no contabamos con la información acustica por lo que se generaron modelos carentes de información contextual. En estos primeros modelos se pudo apreciar una calidad mucho peor en los audios generados, sonando estos sumamente metalicos y carentes de prosodia. Tras un par de iteraciónes y tras agregar los factores contextuales pudimos comprobar que ahora las voces sonaban mucho mas humanas.

HTS tambien brinda la posibilidad de realizar speaker-adaptive training. Esta tecnica permite aproximar un HMM ya entrenado a una voz distinta de la que solo se posee un corpus reducido.

2.1.3. HTS_engine

Finalmente para generar voces con acento extrangero se utilizó hts_engine. Esta herramienta permite interpolar con pesos arbitrarios entre varios modelos para producir un nuevo modelo con una mezcla de la carga fonetica de ambos hablantes y sintetizar audios. Esto nos brinda un gran rango explorativo para experimentar y ajustar la carga fonetica de los modelos originales para acercarnos al objetivo.

Este metodo tiene a su vez cierto sustento teorico en la manera real en la que un no nativo aprende un idioma con una carga fonetica diferente al nativo. Citando un extracto del trabajo *Transcription of Spanish and Spanish-Influenced English, Brian Goldstein, Temple University*:



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/, /r/, /l/, and /d/.

Visto desde nuestra perspectiva, una persona que aprende una nueva lengua, 'interpola' entre aquellos fonemas conocidos y los fonemas 'objetivo' de la nueva lengua.

De esta misma manera pero de manera cintetica buscaremos interpolar entre modelos de distintos idiomas buscando de manera artificial ese mismo comportamiento.

2.2. Entrenamiento

En esta sección presentaremos la metodología utilizada para el entrenamiento de HMMs y la interpolación entre los mismos.

A modo de resumen, estos serán los pasos a realizar:

1. A partir de tres corpus de datos, dos de ellos en castellano y uno en ingles, se realizará un etiquetado fonético de los corpus para su posterior utilización en el entrenamiento de los HMMs.

2. Realizar el entrenamiento de los sistemas (Uno por cada corpus disponible). Para esto contaremos con el framework de modelado de HMMs HTS.

3. Una vez generados los HMMs utilizaremos herramientas provistas por HTS para interpolar entre ellos y así obtener distintos grados de fonética y prosodia inglesa a la hora de sintetizar audios.

Dado que el castellano y el ingles no utilizan los mismos símbolos fonéticos, si queremos sintetizar audios en castellano con el HMM generado con el corpus en ingles, un desafío que deberemos resolver es el de cubrir todos los símbolos fonéticos del castellano por alguno del ingles.

2.2.1. Preparación De los datos

Como ya adelantamos, en este trabajo contamos con tres corpus de datos disponibles:

- *secyt-mujer*: 741 oraciones, 48 minutos de habla en castellano.
- *loc1_pal*: 1593 oraciones, 2 horas y 26 minutos de habla en castellano.
- *CMU-ARCTIC-SLT*: 1132 oraciones, 56 minutos de habla en ingles.

Para los tres corpus se contaba ademas con sus transcripciones grafemicas.

En los inicios del trabajo contabamos con un solo corpus de datos *secyt-mujer* compuesto por 741 oraciones equivalentes a 48 minutos de habla. Para el mismo tambien contabamos con sus transcripciones foneticas y grafemicas anotadas de manera manual.

En primera instancia, realizamos varias pruebas de concepto utlizando HTS y este corpus. Para ello fue neceasrio construir los utternaces del mismo.

Para obtener los utternaces se plantearon varias estrategias posibles. La primera consistió en utilizar alineamiento automatico utilizando EHMM alignment [4]. Los resultados preliminares fueron bastante adversos: los audios generados resultaban poco inteligibles notandose claros defectos acusticos, El mas notable siendo el fonema /rr que se asemejaba mas a una /r.

Utilizando Praat para visualizar el alineamiento entre utternaces y audios, descubrimos que la alineación estaba desfazada. Sospechamos que esto se debió a algun problema con la normalización de los audios.

Dado que para este corpus contabamos con las transcripciones foneticas anotadas de manera manual se procedió a implementar un hibrido con EHMM. De esta manera buscamos mejorar la alineación pero manteniendo el repertorio fonetico, y la metainformación brindada por el alineamiento automatico.

El modelo generado con estos utternaces mixtos resulto ser superior a los generados solo con alineamniento automatico. Aún así los audios sintetizados todavía no alcanzaban una calidad aceptable, sonando metalicos y aguardentosos.

Se pudieron percibir de manera informal otros detalles tales como que la voz original tenía un pitch mayor que la producida por los modelos, al rededor de un 10 %.

En este momento del trabajo obtenemos otro corpus de datos *loc1_pal* con 1593 oraciones en castellano rioplatense que con aproximadamente 2 horas y 26 minutos de habla.

Para este corpus no cotabamos con transcripciones foneticas manuales por lo que nos vimos forzados a utilizar EHMM nuevamente. Aún así, los resultados fueron muy superiores a los conseguidos con *secyt-mujer*. Los audios sintetizados resultaban inteligibles y con un marcado acento rioplatense. Tras algunas pruebas de concepto donde se experimentó con varios valores de GAMMA, rango de las frecuencias principales, y otros parametros que concideramos podían afectar la calidad de la voz decidimos que la voz ya había alcanzado la calidad adecuada para proseguir con 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. Concideramos que esto juega un papel predominante en la calidad de los TTS generados, aún cuando se utiliza un metodo de etiquetado puramente automatico y propenso a errores en el alineamiento.

Por ultimo utilizamos el corpus *CMU-ARCTIC-SLT* con 1132 oraciones y 56 minutos de habla, disponible en la pagina de hts [5].

Para este trabajo todos los audios usarán sampling rate de 48kHz, precisión de 16 bits, mono.

El rango de extraccion de frecuencia principal para utilizado fue de 100 hz a 350hz.

Una lista extensiva de los parametros utilizados para el entrenamiento se puede ver en el apendice 3

2.2.2. Repertorio Fonetico y Mapeo De Fonemas

Para las transcripciones foneticas, tanto de los audios en ingles como en castellano, utilizamos los repertorios foneticos brindados por festvox (ver apendice 1).

El primer desafío que se presenta es que estos repertorios foneticos no tienen un mapeo directo con el Alfaveto Fonetico Internacional: por ejemplo con este repertorio fonetico, en el castellano existen tres fonemas distintos para la /i/. Consideramos que esta decición por parte de festvox proviene de la necesidad de poder diferenciar la /i/ acentuada de la no acentuada y de aquella presente en los diptongos.

Por otra parte, surge aquí un problema: como sintetizar oraciones en castellano utilizando un repertorio fonetico distinto, donde incluso la cantidad de fonemas es diferente. Como solución a esto desarrollamos de manera perceptual e iterativa, un mapeo del ingles al castellano en el que cubriremos cada fonema del castellano por al menos uno del ingles. El mapeo que concideramos devolvió los mejores resultados puede observarse en el apendice.

Los fonemas marcados como notUsed los consideramos lo suficientemente diferentes como para no mapearse a nigun fonema del castellano.

Ademas para completar el repertorio, se tomaron la mitad de los fonemas /r/ y se remplazaron con /rr/ y de manera similar se tomaron la mitad de los fonemas etiquetados como /hh/ y se remplazaron con /g/.

Utilizamos este mapeo para generar un tts en ingles capaz de sintetizar oraciones en castellano. Por supuesto los resultados obtenidos sintetizando audios de esta manera generan audios incomprensibles y de muy baja calidad.

En el proximo paso procederemos a realizar mezclas entre el tts en castellano y el tts presentado aquí para generar un nuevo tts donde se pueda hacer un ajuste gradual de cada uno de estos modelos.

2.2.3. Interpolación entre modelos

Una vez generados ambos modelos con fonemas 'similares' proseguimos realizando pequeñas pruebas internas para probar la efectividad del metodo y concluimos que eran satisfactorias, fue posible generar oraciones donde la carga fonetica podía reconocerse como estadounidense (detalles distintivos como la /r mas suavizada, o las vocales mas abiertas).

2.2.4. Speaker-adaptive Training

Se realizaron varias pruebas de concepto donde habiendo entrenado un HMM con *CMU-ARCTIC-SLT* se le realiza speaker adaptation con el corpus de *loc1_pal*, pero resultaron no concluyentes. El HMM final perdía completamente las caracteristicas y el acento de *CMU-ARCTIC-SLT* por lo que se decidió no proseguir por este camino.

3. Experimentación

En la siguiente aparatado intentaremos validar dos hipótesis: que el modelo generado realmente puede ser identificado como un hablante extranjero de habla inglesa y al mismo tiempo que este posee un grado de inteligibilidad aceptable.

Para eso se condujo una encuesta perceptual donde, dado un participante, se le presentó un audio con una oración semánticamente impredecible, fonéticamente balanceada y con distintos grados de mezcla de español e ingles, se le pidió que la transcribiera y que intentara identificar la nacionalidad del mismo.

Para la experimentación, se generaron diez oraciones distintas variando el nivel de mezcla de los modelos generados entre 30 % de ingles - 70 % castellano hasta 70 % de ingles - 30 % de castellano.

Los utternaces utilizados fueron generados de manera algorítmica tomando palabras de manera aleatoria de una lista de sustantivos, adjetivos y verbos y realizando correcciones donde fuera necesario para mantener el balanceo fonético.

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

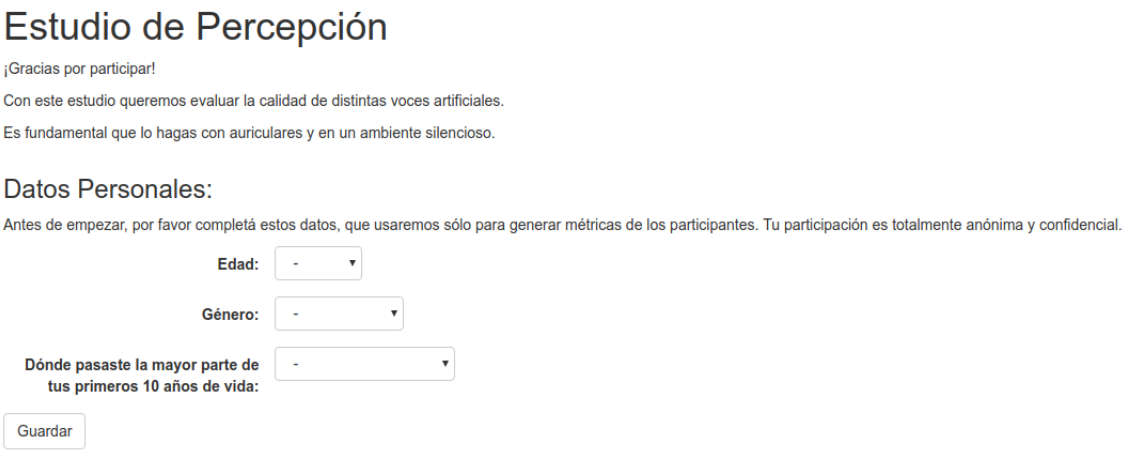
La encuesta se realizó a través de internet, con el mismo set de instrucciones para todos los participantes y pidiendo como requerimiento la utilización de auriculares. Cada participante podía contestar un máximo 5 veces (otorgándoles siempre audios distintos).

La misma se llevó a cavo desde el 18 de octubre de 2017 hasta el primero de diciembre del mismo año, tiempo durante el cual fue publicada en distintas redes sociales y listas de emails de la facultad.

3.1. Interface

A continuación se presenta la interfase que utilizamos para realizar la encuesta junto con las decisiones de diseño que fueron tomadas a lo largo de la misma.

Al entrar en la pagina de la encuesta todos los participantes fueron presentados con la siguiente pantalla principal:

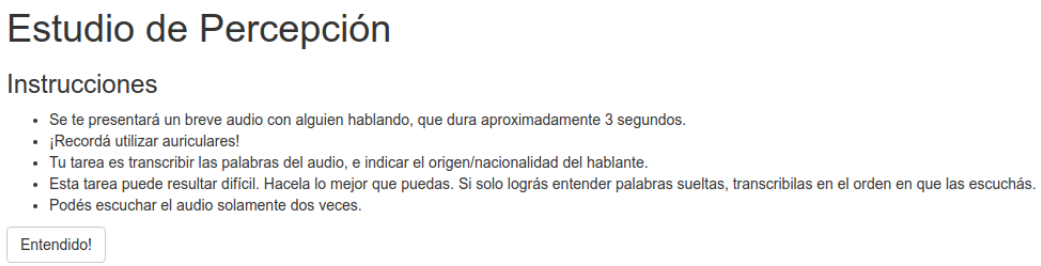


Con el objetivo de no influir el las respuestas de los participantes participantes, se procuró darles a los participantes la información indispensable para completar la encuesta. En ningún momento se especifica que buscábamos con este estudio.

Con la intención de que los resultados fueran lo menos variables posibles, se le pidió a cada participante que realizara la encuesta con auriculares y en un lugar silencioso.

Además, le pedimos a cada participante que indique su genero, su edad y la provincia en la que transcurrido la mayor parte de su infancia.

Una vez que completados estos datos, se les presentaba otra vista con las instrucciones especificas para completar la encuesta:



Una vez presionado el botón de “entendido!” se les presentaba un audio, que podían escuchar un máximo de 2 veces, una caja de texto libre donde plasmar la transcripción del mismo y una caja de texto libre donde podían escribir la nacionalidad correspondiente a la voz.

Una vez que la respuesta era guardada, si el participante todavía no había contestado 5 veces, se le preguntaba si quería continuar transcribiendo otro audio, o de caso contrario se le presentaba un mensaje donde se le indicaba que ya podía cerrar la encuesta.

3.2. Resultados

A modo de introducción, comenzaremos mostrando los datos demográficos de los participantes. Continuaremos con un análisis mas exhaustivo de la inteligibilidad y por otro lado de la nacionalidad atribuida. Por ultimo para contestar compondremos estos dos ejes para intentar contestar la hipótesis original.

3.2.1. Datos demográficos

Se encuestaron 109 participantes de los cuales se obtuvieron 352 resultados.

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:

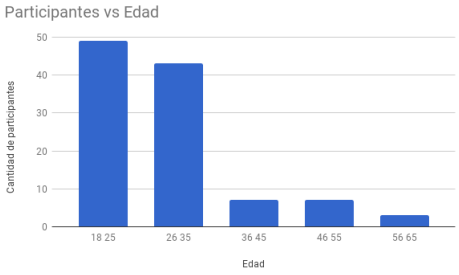


Figura 4: Edad de los participantes

Con respecto al genero de los participantes, 187 respuestas fueron brindadas por participantes del genero femenino mientras que 163respuestas fueron brindadas por participante del genero masculino.

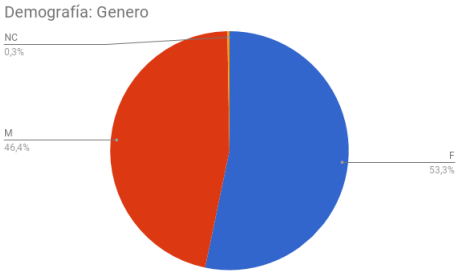


Figura 5: Genero

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

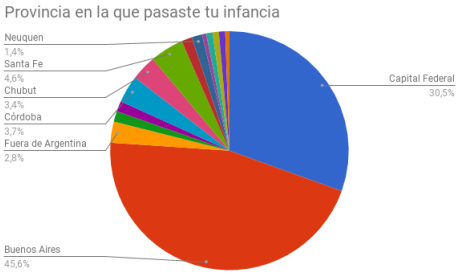


Figura 6: Figure caption

3.2.2. Inteligibilidad

Para el análisis de resultados utilizaremos la distancia de Levenshtein con inserciones, remociones y reemplazos. Respetando los acentos pero sin tener en cuenta mayúsculas o minúsculas.

Presentamos aquí los resultados obtenidos sin ningún tipo de modificación:

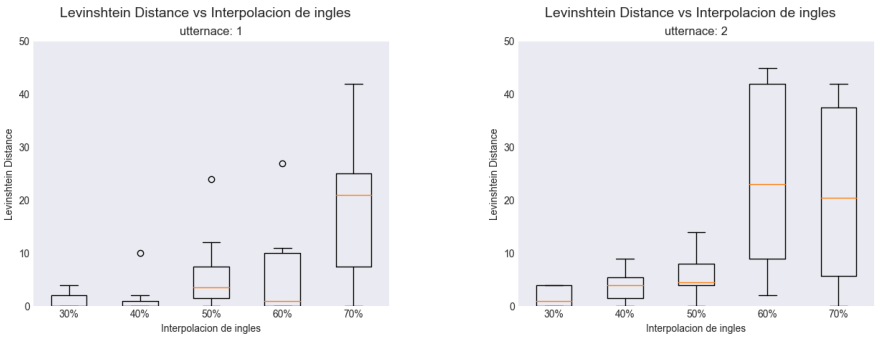


Figura 7: Utternace 1 y 2

Como puede verse en la mayoría de los utternaces se puede observar que hasta el 50 % de mezcla castellano-ingles, se conserva el un buen grado de inteligibilidad, rondando la distancia de Levenshtein al rededor de 10 a 20 caracteres. Pasados el 60 % de ingles, se observa una disminución brusca en la inteligibilidad, llegando a una distancia de 45 caracteres.

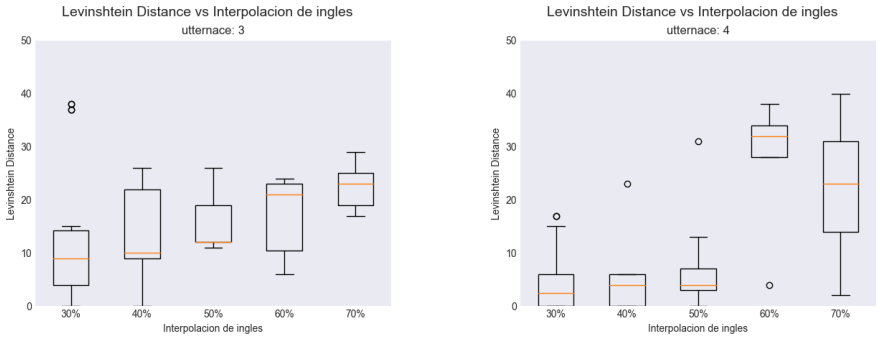


Figura 8: Utternace 3 y 4

Analizando detenidamente las transcripciones obtenidas observamos algunas fallas sistemáticas que podrían generar ruido en el análisis, tales como:

- Los participantes escribieron de manera diferente cuando no entendieron un segmento del audio.
  - Muchos de ellos escribieron: "...", "...." o simplemente omitieron la palabra.
  - En casos menos comunes: "***", "???", "blablabla".
- En casos donde no comprendieron ninguna palabra del audio escribieron cosas como "no entendí nada", "nada", dejaron el campo vacío, etc.
- Utilización de signos de puntuación en las oraciones:
  - Puntos finales para expresar el final de la oración o expresiones como "(?)".
  - En un caso extremo, un participante el participante transcribió " "tu estrecho posavasos", grito la fechoría", cuando el Utternace original solo decía "tu estrecho portavasos gritó la fechoría".
- Omisión de acentos y faltas ortográficas en palabras que no ambiguas.Ejemplo: "grunion" en vez de "gruñón"

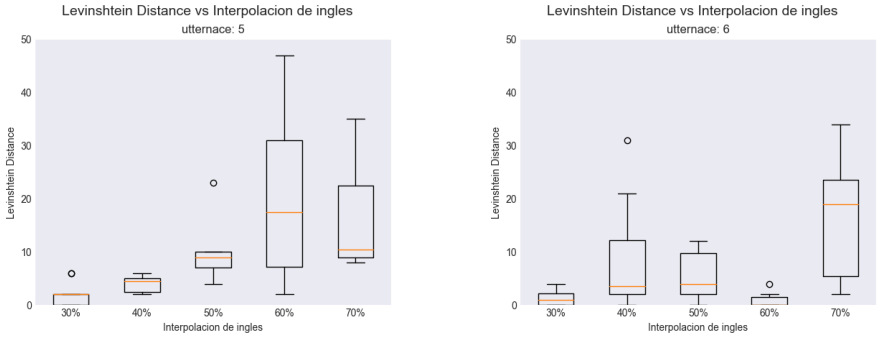


Figura 9: Utternace 5 y 6

Para disminuir ruido de la muestra, se decidió realizar una limpieza de los datos donde consideramos que no era disruptiva.
Los cambios fueron:

- corregir "ni" por "ñ " en la palabra grunion.
- Remoción de todos los signos de puntuación.
- Remoción de expresiones como "blabla", "no entendí.º cualquier otra que exprese ininteligibilidad de una palabra u oración.
- Corrección de acentos en palabras no ambiguas: "botón", "prefirió", "recorrió", "chupetín", "riñón", "gruñón".

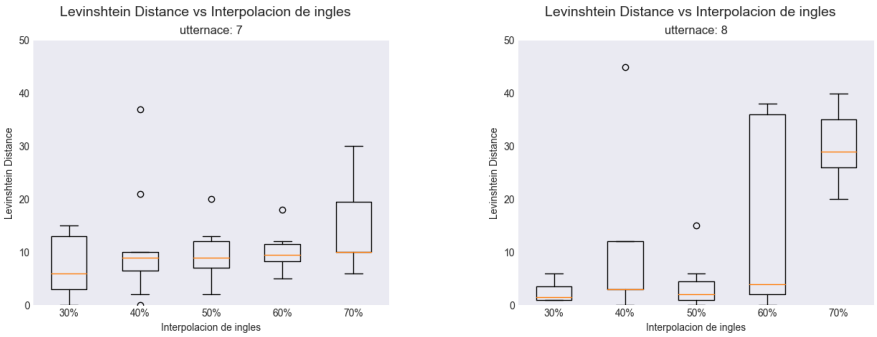


Figura 10: Utternace 7 y 8

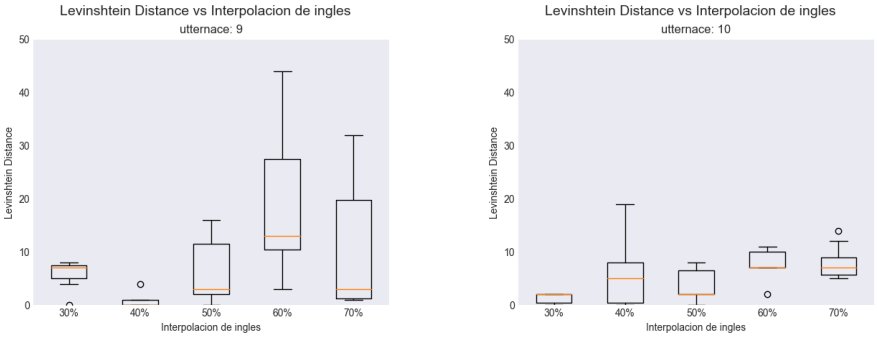


Figura 11: Utternace 9 y 10

Aquellas palabras que presentan ambivalencia, como : “concluyó” no fueron modificadas ya que concluyó/concluyo son validas.

Esta normalización no solamente ayudará a disminuir la propagación de los datos sino que nos permitirá entender de manera mas intuitiva que significa la distancia de Levenshtein en cada caso.

3.2.3. Datos normalizados

Con esta estandarización de los datos, trataremos de darles un peso intuitivo que nos permitan sistematizar el análisis.

Por ejemplo, tomando el Utternace 8 de las frases utilizadas en la experimentación:

- "Las acongojadas cotorras sonrieron a mi círculo"

Podemos observar que:

- 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: ""

A partir de esto, para este apartado vamos a tomar que una distancia de Levenshtein

- Distancia 0-10: Buena Inteligibilidad
- Distancia 10-20: Mediana Inteligibilidad
- Distancia 20-40: Baja Inteligibilidad
- Distancia 40-: Inteligibilidad nula

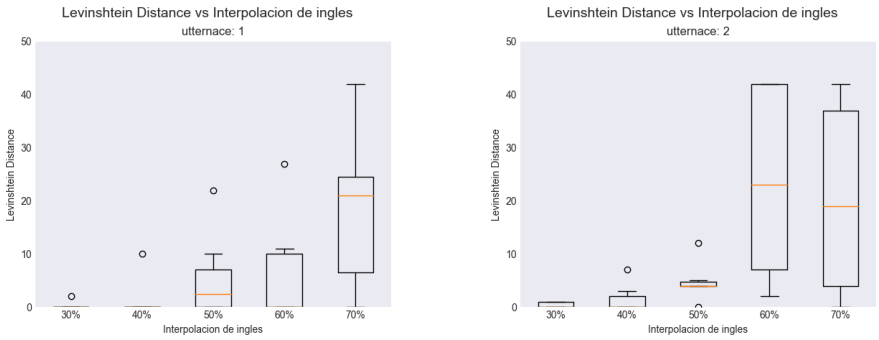


Figura 12: Utternace 1 y 2 Normalizados

Con este baseline, podemos ver que para una interpolación de ingles de 30 %, 96 de los 106 participantes comprendieron de manera adecuada el texto con una inteligibilidad alta. Los 10 restantes obtuvieron una inteligibilidad media.

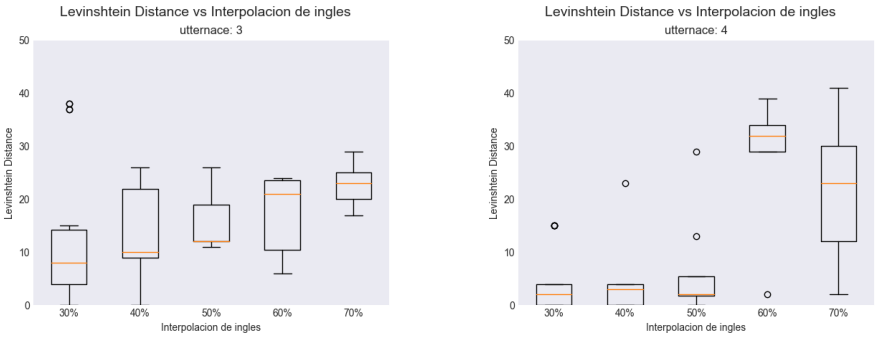


Figura 13: Utternace 3 y 4 Normalizados

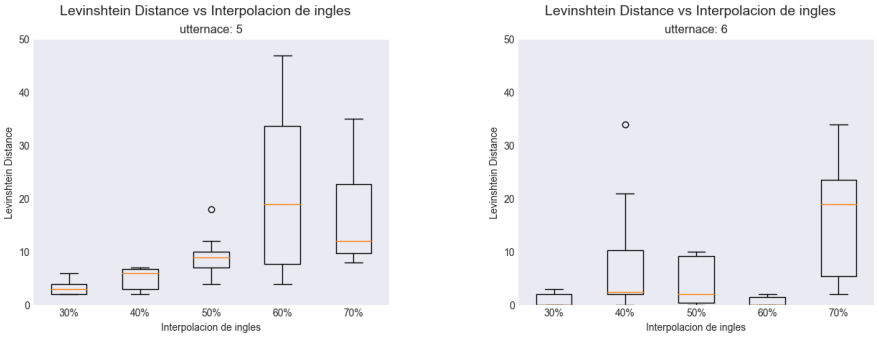


Figura 14: Utternace 5 y 6 Normalizados

Para la interpolación 40 % ingles - 60 % castellano, de un total de 67 participantes, 57 obtuvieron una inteligibilidad alta, 2 una inteligibilidad media, 5 una baja y 3 una inteligibilidad nula.

Para la interpolación 50 % ingles - 50 % castellano, de un total de 75 participantes, 57 obtuvieron transcribir el audio demostrando una buena inteligibilidad, mientras que 14 tuvieron una inteligibilidad media y 4 una inteligibilidad baja.

Para todas las interpolaciones enunciadas previamente, los errores mas comunes varían desde falta de acentos en palabras como "concluyó" ambiguas hasta faltas de inteligibilidad en palabras con cierta complejidad como "aguileña" o "gruñón".

Para el Utternace 3: "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", y "enjoyado" que transcribieron como "enfollado", "enrollado" y la conjugación exacta del verbo comprar.

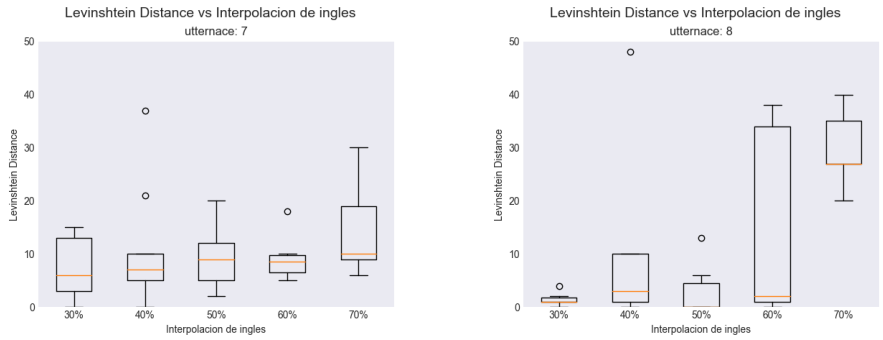


Figura 15: Utternace 7 y 8 Normalizados

Para los grados de interpolación 60 % ingles - 40 % castellano y 70 % ingles - 30 % castellano, se pueden observar un aumento notable de la variabilidad en las respuestas. Para el primero, de las 70 respuestas obtenidas, 40 participantes lograron transcribir con un buen grado de inteligibilidad los audios, 6 obtuvieron una inteligibilidad media, y 24 transcribieron el audio con una inteligibilidad baja o nula.

Para 70 % ingles - 30 % castellano, la diferencia es todavía mas marcada, de los 68 resultados obtenidos, 28 lograron transcribir el audio con un buen grado de inteligibilidad, 8 con un grado medio y 32 con un grado bajo o nulo de inteligibilidad.

Consideramos estos resultados tan dispares pueden deberse a dos motivos:

El primero, características particulares de los participantes y sus capacidades para discernir palabras incluso cuando presentan defectos en la pronunciación del hablante. En particular, el Utternace 4 muestra como para un mismo audio, con características similares 2 participantes de los 9 que realizaron la transcripción, obtuvieron distancias 2 y 6 en sus transcripciones.

El segundo motivo puede deberse a características particulares de los utternaces o del modelo utilizado para generar la voz: el Utternace 10, donde el 6

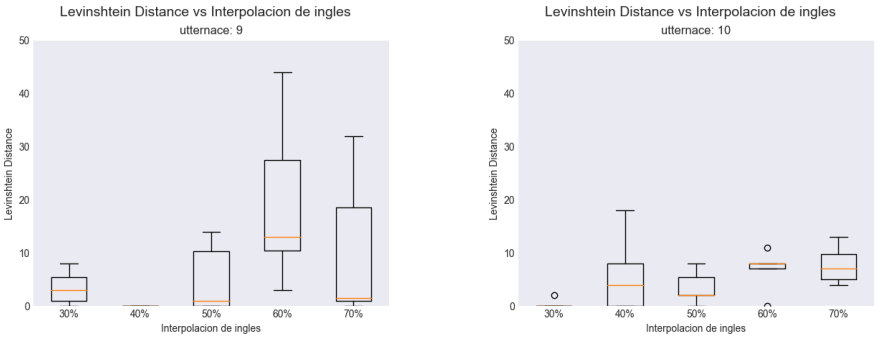


Figura 16: Utternace 9 y 10 Normalizados

de los 8 participantes obtuvieron una buena transcripción del audio, y el Utternace 8, donde todos los participantes transcribieron el audio con inteligibilidad baja o nula, parecen demostrar esto. O bien la dificultad de los utternaces es variable o, lo que es todavía mas probable, llegado cierto punto en la interpolación, algunos fonemas empiezan a "romperse" o se alejan demasiado del fonema castellano correcto y terminan por disminuir la claridad de la voz.

3.2.4. Análisis de Nacionalidad

En esta sección analizaremos los resultados de las nacionalidades que los participantes atribuyeron a la voz.

Dado que en esta instancia se le le permitió a los participantes ingresar texto libre las respuestas resultaron bastante heterogéneas. Los participantes tomaron la consigna de manera diferente, pudiendo encontrarse respuestas que no pueden ser atribuidos exactamente a una nacionalidad. Ejemplo de algunas respuestas: Latino, Anglo, Robot, España (sur).

Consideramos que las respuestas de la índole "robot", "es una voz artificial", no son validas ya que no aportan información para esta investigación.

Por esta razón, en esta instancia decidimos agrupar las respuestas en cuatro grupos lógicos:

- Hispanohablante: Latino, Argentino, Español, Uruguayo, Centroamericano, Boliviano, Mexicano, Colombiano
- Angloparlante: Estadounidense, Ingles, Irlandés, Canadiense, Anglo
- No sabe/No contesta: Robot, no se
- Otro: Ruso, Brasiltiño

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

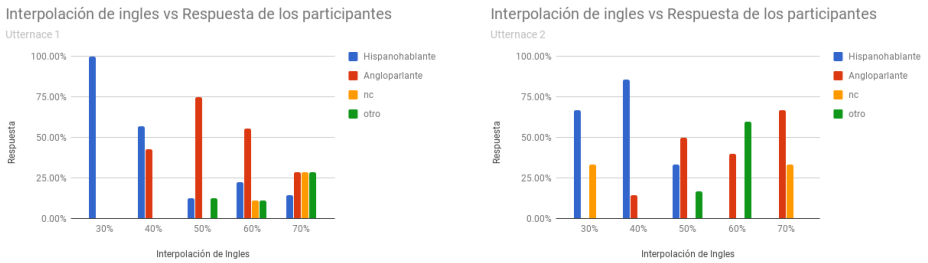


Figura 17: Utternace 1 y 2

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

Para una mezcla de 40 % ingles, puede verse que no hay una decisión concluyente con respecto a la nacionalidad del hablante. Por ejemplo en el Utternace 3 el 80 % de los participantes coincide que la voz pertenece a un hablante de habla hispana, mientras que en el Utternace 9 el 60,00 % de los participantes considera que la voz pertenece a un hablante de habla anglosajona.

Esta gran disparidad de resultados entre distintos utternaces se puede atribuir a las características particulares de cada Utternace. En particular el Utternace 9: "Ese gruñón perro prometió a esos cuñados" contiene una /r vibrante que resulta muy notoria al pronunciarse con una intensidad menor a la esperada y es atribuida, en general, a un hablante extranjero.

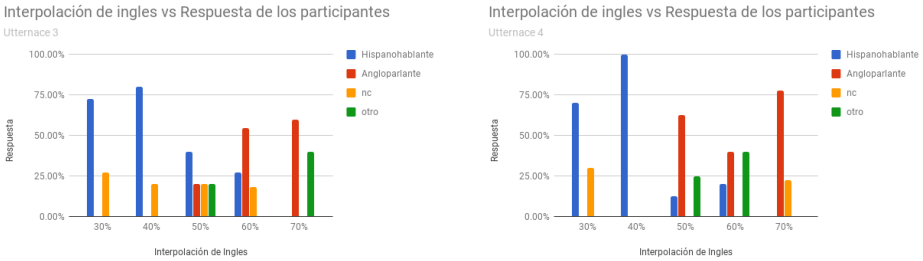


Figura 18: Utternace 3 y 4

Bajo esta suposición observamos que los otros utternaces que presentan este fonema:

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

También presentan un mayor porcentaje de atribuciones a nacionalidad anglosajona.

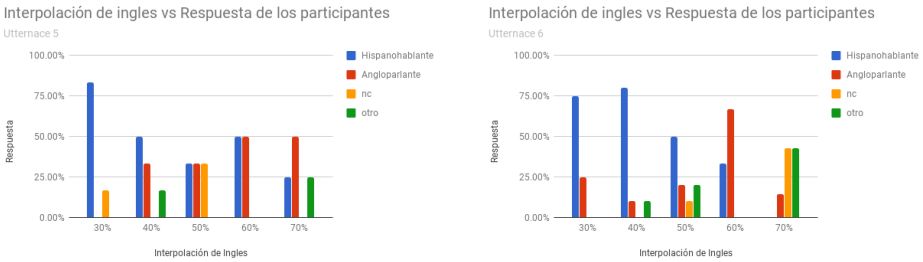


Figura 19: Utternace 5 y 6

Con 50 % y 60 % de ingles los resultados son similares. obtenemos que aproximadamente en el 50 % de los utternaces, mas de la mitad de los participantes consideraron que la voz pertenecía a un anglosajón hablando castellano. Para estos grados de interpolación también podemos observar que en un 80 % de los utternaces al menos un 20 % de los participantes atribuyen la nacionalidad del hablante a un no nativo no anglosajón hablando castellano.

Con 70 % de interpolación, en el 80 % de los utternaces se puede apreciar que al menos 50 % de los participantes dijo que el hablante era de origen anglosajón. Mas aún, en el 40 % de los utternaces el 75 % de los participantes coincidió que la voz era de angloparlante. También podemos ver que para este grado de interpolación en el 70 % de los utternaces ningún participante considera que la

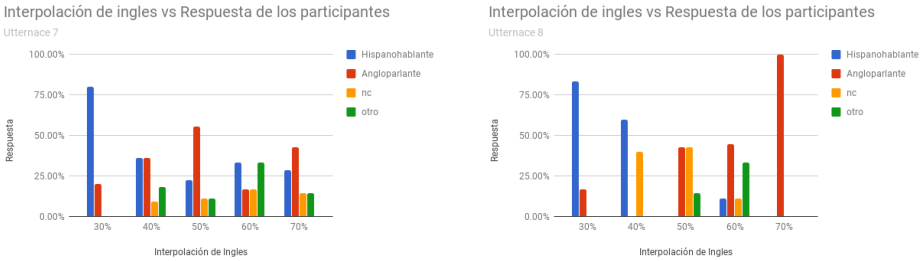


Figura 20: Utternace 7 y 8

voz sea de habla hispana. En el 30 % restante, 25 % de los participantes o menos concideran que la voz pertenezca a un hispanohablante.

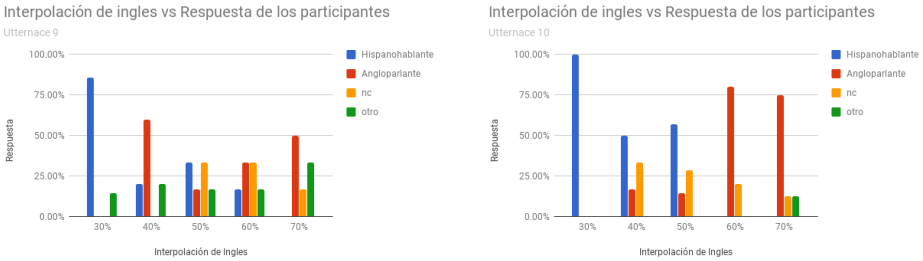


Figura 21: Utternace 9 y 10

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

3.2.5. Resultados Generales de la experimentación

Por ultimo visualizamos mostraremos la distancia de Levenshtein superpuesto con con la probabilidad de un participante de reconocer la voz como un hablante anglosajón.

Volviendo a la hipótesis original, podemos ver que esta técnica permite generar una voz que pueda ser identificada como un extranjero hablando ingles, con un grado de efectividad que varía desde el 60 % hasta el 100 % dependiendo del Utternace elegido y el grado de interpolación.

También es interesante observar casos como el que se presenta comparando el Utternace 10 y el Utternace 8, ambos con 70 % de mezcla de ingles, que en cuanto a inteligibilidad se encuentran en extremos opuestos, muestran que aproximadamente un 80 % y un 100 % de participantes identificaron como nativo anglosajón. Esto nos da a pensar que la inteligibilidad de una oración y su probabilidad de ser identificado como un hablante ingles son variables independientes y que este ultimo factor este mas ligado a otros factores como la

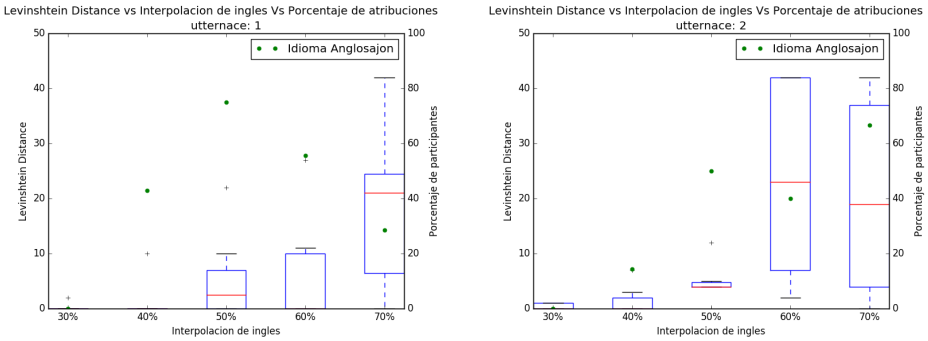


Figura 22: Utternace 1 y 2

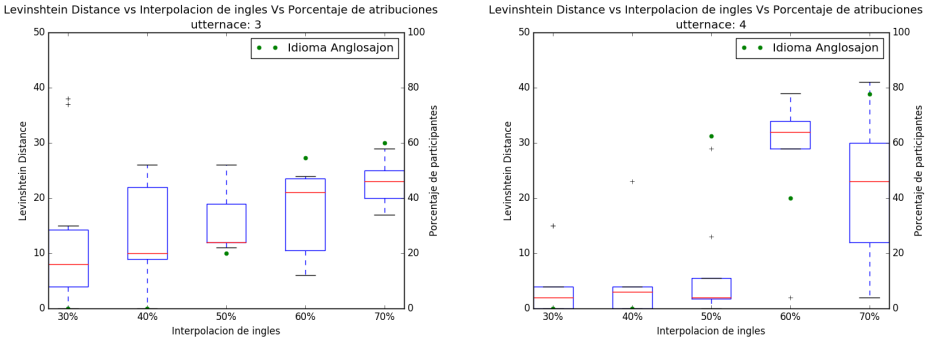


Figura 23: Utternace 3 y 4

sonoridad de ciertos fonemas o la prosodia general de la voz.

Este no es un caso aislado, véase que lo mismo sucede con el Utternace 4 y 6 con 60 % de mezcla de ingles, si bien las inteligibilidades están en extremos opuestos, sus probabilidades de ser identificados como hablantes extranjeros difieren en menos del 20 %.

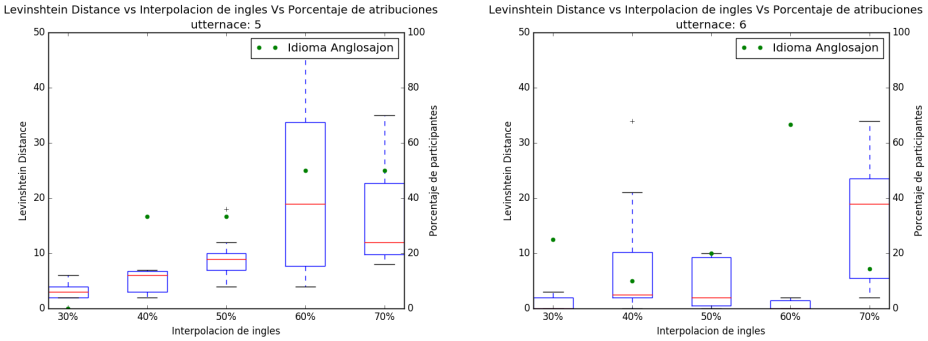


Figura 24: Utternace 5 y 6

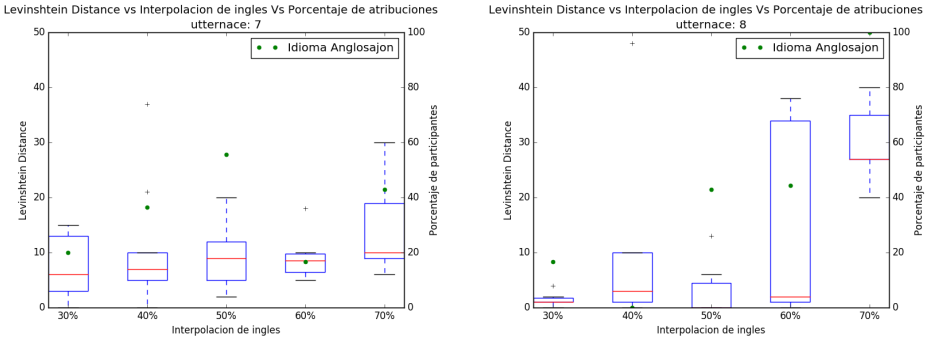


Figura 25: Utternace 7 y 8

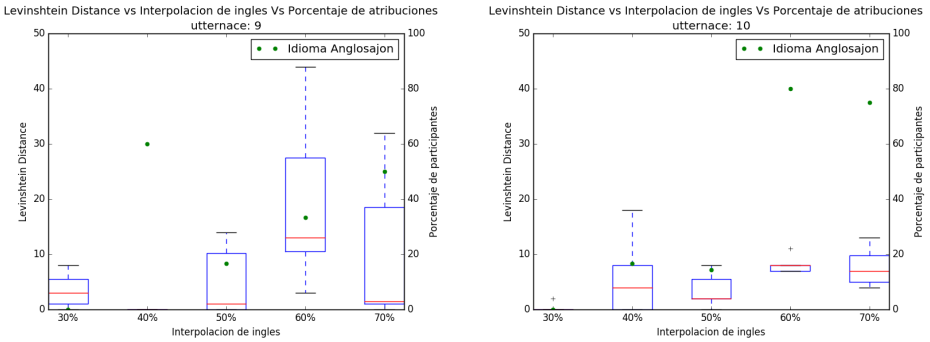


Figura 26: Utternace 9 y 10

4. Trabajo Futuro

Como fue discutido en la sección de experimentación, una interpolación general puede producir que ciertos fonemas se alejen demasiado del fonema real del castellano, disminuyendo la inteligibilidad de la voz sintetizada. Un posible camino a seguir es realizar una interpolación controlada que permita regular cada fonema por separado. Para fonemas que puedan resultar problematicos como el caso de la /r bibrante el grado de interpolación podría dejarse mas cercano al castellano, mientras que para fonemas con comportamientos mas similares el grado de interpolación podría llevarse mas cerca del modelo ingles.

5. Apendice

5.1. Lista de Fonemas

Castellano	Ingles
a	aa
a1	ae
b	ah
ch	ao
d	aw
e	ax
e1	ay
f	b
g	ch
i	d
i0	s
i1	dh
k	eh
l	er
ll	ey
m	f
n	g
ny	hh
o	ih
o1	iy
p	jh
r	k
rr	l
s	m
t	n
u	ng
u0	ow
u1	oy
x	p
-	r
-	sh
-	t
-	th
-	uh
-	uw
-	v
-	w
-	y
-	z
-	zh

5.2. Mapeo Fonemas del Ingles-Castellano

Ingles	Castellano
aa	a1
ae	a
ao	o
ou	o1
b	b
ch	ch
d	d
dh	d
dx	dx
eh	e
el	e1
em	em
en	en
er	er
ei	ei
f	f
g	notUsed3
hh	h
hv	hv
ih	i1
iy	i
k	k
l	l
m	m
n	n
nx	n
ng	ng
p	p
r	r
s	s
t	t
uh	u1
uw	u
w	u0
th	notUsed
v	v
jh	ll
y	y
sh	sh
zh	zh
z	notUsed2

5.3. Parametros utilizados para el entrenamiento

6. Referencias

- [1]Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization, Heiga Zen, Member, IEEE, Norbert Braunschweiler, Sabine Buchholz, Mark J. F. Gales, Fellow, IEEE, Kate Knill, Member, IEEE, Sacha Krstulovic, and Javier Latorre, Member, IEEE
- [2]Speaker Similarity Evaluation Of Foreign-accented Speech Synthesis Using Hmm-based Speaker Adaptation, Mirjam Wester, Reima Karhila
- [3]Simultaneous Modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura
- [4] SUB-PHONETIC MODELING FOR CAPTURING PRONUNCIATION VARIATIONS FOR CONVERSATIONAL SPEECH SYNTHESIS, Kishore Prahallad, Alan W Black and Ravishankhar Mosur https://www.cs.cmu.edu/ awb/papers/ICAS
- [5] http://hts.sp.nitech.ac.jp/?Download