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

Abstract

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 mi persona favorita: la transformada de Fourier

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, traducción automática, etc.
En las últimas décadas se han visto grandes progresos en este campo. sAlgunos sistemas son capaces de modelar con cierto grado de efectividad cuestiones tales como la el acento, el tono y la entonación de un hablante (es decir, su prosodia). Por otro lado tambien se ha visto un progreso en modelar emociones, o las intenciones discursivas (Cuestiones prosodicas que pueden darnos a entender que una oración es una pregunta, una orden, etc).
Actualmente se considera que el estado del arte para la síntesis de voz es el entrenamiento con redes neuronales profundas (DNN). 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 permita sintetizar habla. Para esto utilizaremos HTS[10], un framework de entrenamiento y sintesis de voz basado en HMM+GMM.
Consideramos que este método, si bien no nos permitirá obtener la mejor voz posible, nos permitirá entender mejor los features con los que trabajamos y por lo tanto modificarlos según consideremos conveniente.
En este trabajo de tesis se estudia una manera posible de generar un TTS basado en HMMs capaz de sintetizar habla en español con acento extranjero. Existen muchos motivos por los que podría construirse un sistema cone 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 poner un ejemplo, al pedir indicaciones de como llegar a algún lugar uno no deposita el mismo nivel de confianza si la persona que responde suena oriundo de la localidad que a alguien que suena extranjero. Por ultimo 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.
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 [7], donde se describe de manera detallada las desiciones de diseño utilizadas en HTS, así como el modelado de parámetros para la construcción de HMMs, el modelado de cada fonema utilizando Mel-Cepstral, etc.
Como introducción a este trabajo comenzaremos analizando las decisiones de metodología utilizadas a lo largo de la investigación, como así también detalles teóricos como el mapeo de fonemas necesario para adaptar el repertorio fonético del inglés al castellano, etc.
En las siguiente secciones se detallarán distintos temas que fueron necesarios abordar para llevar a cabo esta tesis. En orden de aparición estos son:

- Comenzando en la sección 2.1 se presentan las técnicas utilizadas para el etiquetado

fonético de los distintos corpus sobre los que trabajaremos.

- En la sección 2.2 se presentará un mapeo entre los fonos del castellano y los del inglés.
- En la sección 2.3 y 2.4 presentamos las herramientas provistas por HTS para combinar modelos y poder sintetizar habla con distintos grados de mezcla fonética y prosódica de inglés-castellano.
- En la sección 2.5 se detallaran aspectos técnicos del trabajo, así como especificaciones de como modela el audio HTS, estructuras utilizadas durante el entrenamiento y otras herramientas utilizadas.
- Por ultimo, en todo el capitulo 3, intentaremos validar que el sistema construido realmente cumple con las caracteristicas deseadas, exponiendose en la sección 3.1 los aspectos metodologicos de la evaluación y en las subsiguientes secciones 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 el capitulo 2.5.

2.1.1. Primer corpus en castellano

Al inicio de la investigación comenzamos con un solo corpus de datos, *secyt-mujer*[1]. El mismo está compuesto por 741 oraciones cortas declarativas habladas por una locutora profesional en español 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 principalmente en una lista de fonos dónde se indica cuando 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: como suena un fonema cuando está seguido de algún otro, al principio de una oración, si se encuentra en un diptongo). 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 [8] empleando Festival y Festvox. Estos programas tienen como ventaja que tanto la anotación del corpus como la generación de oraciones utilizadas para la síntesis presentan el mismo repertorio fonético. Esto es útil ya que podemos garantizar que el entrenamiento y la síntesis están utilizando el mismo método de generación de trazas/oraciones.

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 Praat para visualizar el alineamiento entre las transcripciones fonéticas obtenidas y los audios, descubrimos que la alineación estaba desfasada algunas milésimas de segundo. 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 estaban demasiado alejados de los audios objetivo.

Dado que para el corpus contamos con las transcripciones fonéticas anotadas de manera manual, procedimos a implementar un híbrido con EHMM. Este híbrido consistirá en

tomar las anotaciones hechas a mano para cada fonema por un lado y la información contextual y el repertorio fonético generado a partir del proceso de EHMM por otro y combinarlos en una nueva transcripción fonética. De esta manera buscamos mejorar la alineación pero manteniendo el mismo repertorio fonético y la misma meta-información brindada por el alineamiento automático.

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 todavía 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.

2.1.2. Segundo Corpus En Castellano

En este punto de la investigación obtenemos un segundo corpus de datos: *loc1_pal* [2] 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 apreciamos mayores desfasajes con las anotaciones fonéticas.

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[5]. 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 *CMU-ARCTIC-SLT*[3] con 1132 oraciones en inglés y 56 minutos de habla articuladas por una mujer estadounidense, disponible en la página de HTS [10]. Ya que este corpus venía a modo de demo con HTS, asumimos que los parámetros de entrenamiento y las transcripciones fonéticas ya habían sido seleccionadas de manera apropiada, por lo que no intentamos mejorar la calidad de las transcripciones fonéticas mas allá de lo que la demo ofrecía.

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 mas detenidamente los aspectos técnicos de los corpus utilizados.

2.2. Repertorio Fonético y Mapeo de Fonemas

Como ya adelantamos brevemente en el apartado anterior, para la generación de oraciones para la síntesis utilizamos Festvox y Festival. Al igual que con las transcripciones fonéticas, estas herramientas nos permitirán traducir oraciones grafémicas (texto) a una lista de fonos y metadata utilizados por HTK para sintetizar audio.

El primer desafío que se presenta es que estos repertorios fonéticos no tienen un mapeo directo entre el inglés y el 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 e ingles.

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

De esto, surge el siguiente problema: cómo sintetizar oraciones en castellano utilizando un repertorio fonético en inglés, donde incluso la cantidad de fonos es diferente. La manera que encontramos de abordar esto fue confeccionar una tabla donde cada fono del castellano estuviera mapeado a uno del inglés. En otras palabras, antes del entrenamiento del modelo, tomamos los fonos generados por festival para el corpus en inglés y los reemplazamos con fonos del castellano de la manera ilustrada en la tabla 2.

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	-

Tab. 2.2: Mapeo Fonético

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 fonos a partir de otros, 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 al suyo. Citando un extracto del trabajo *Transcription of Spanish and Spanish-Influenced English, Brian Goldstein, Temple University*[9]:

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 ingles (glad). Notamos este error recién al hacer la evaluación perceptual, como se describe en la sección 3.6

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 exacto, mapenado por ejemplo el fono /w/ (*twentieth*) del ingles al fono /u0/ (*ahuyentar*) del castellano o el fono /uw/ (*two*) por el fono /u/ (*cumplió*).

Un mapeo que podría resultar controversial es del fono /jh/ (*danger*) del ingles por el /ll/ (*billete*) del castellano. Si bien esta decisión no puede resultar ser la mejor, por ejemplo /ll/ podría ser mapeado a /sh/ (*ash*), consideramos que para este primer trabajo resulta suficientemente buena.

De manera similar el inglés carece del fono vibrante múltiple alveolar sordo /r/ (*perro*) y dado que el fono /ɾ/(*pero*) ya estaba siendo utilizado, no podíamos realizar un mapeo tan directo. Como solución tomamos la mitad de los fonos /ɾ/ y los reemplazamos con /r/.

Aquellos fonos que consideramos suficientemente disímiles del castellano, como es el caso de la /sh/, /z/, etc 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 solamente el corpus en ingles para sintetizar oraciones en castellano, aunque como es de esperar, dado que el corpus de entrenamiento es tan distinto de las oraciones que queremos sintetizar (Por cuestiones como que las combinaciones de fonos del ingles y el castellano son diferentes, las reglas prosódicas y las acentuaciones vocálicas difieren entre los idiomas) los audios sintetizados resultan incomprensibles.

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 por *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 mas detallada del tema ver 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 el apartado anterior, presenta problemas fonéticos graves, haciendo que las oraciones resultaran poco naturales y difíciles o imposibles de comprender. En el medio de la 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/ mas suavizado, o la pronunciación de vocales más abiertas, pero aún conservando cierto grado de inteligibilidad.

Dado que HTS modela de manera conjunta la acústica y la prosodia, 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 que pudimos apreciar es que cuanto mas cercano esta el grado de interpolación al modelo en castellano, las características fonéticas se asemejan mas a la de las oraciones del corpus *loc1_pal*, mientras que de manera análoga, cuanto mas grado de inglés tiene, la voz se asemeja a *CMU-ARCTIC-SLT*. Si bien esto

no es un defecto importante, con el motivo de cambiar el menor numero 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-Adaptative Training sobre uno de los modelos.

2.4. Speaker-Adaptative Training

El Speaker-Adaptative 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 tanto en espacio de almacenamiento, tiempo de grabación y etiquetado, por lo que resulta mas económico generar una nueva voz sintética a partir de un modelo bien generado y adaptándolo luego con características del nuevo corpus.

Nuestro objetivo para este trabajo es utilizar esta herramienta para aproximar las características de uno de los hablantes al otro para que sus identidades fueran indistinguibles.

Como prueba de concepto se tomó el corpus *CMU-ARCTIC-SLT* y se le realizó Speaker Adaptation junto con *loc1_pal* utilizando la demo presente en la sección de descargas de HTS para el entrenamiento.

Dentro del adaptative training existen varias técnicas, en este trabajo utilizaremos offline supervised adaptation, que tiene como requisito adicional conocer los oraciones del segundo corpus.

Las pruebas no resultaron concluyentes, las oraciones sintetizadas no solamente perdían la identidad del hablante original sino también sus características fonéticas. Dicho de otra manera, si lo que buscábamos era que obtener un hablante ingles que pudiera ser reconocido como el mismo locutor que *loc1_pal* pero con sus características fonéticas intactas (una pronunciación suavizada del fonema /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[4] [6], dada la complejidad del método y los largos periodos que son necesarios para entrenar un modelo (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 utilizadas para este trabajo. Así también como los comandos con los que se realizó en preentrenamiento, el entrenamiento y la generación de oraciones para la tesis.

2.5.1. Festival y Festvox: Generación las transcripciones fonéticas

Festival 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 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 utilizaremos Festival y Festvox para generar los oraciones requeridos tanto para el entrenamiento como para la síntesis de audios. Estos consisten básicamente

en una transcripción fonética de los audios dividida en segmentos temporales y datos contextuales tales como la cantidad de silabas en la palabra siendo transcrita, fonemas que preceden y proceden al actual, etc.

A modo de guía a continuación mostraremos como es que utilizamos estas herramientas para generar las transcripciones fonéticas deseadas usando EHMM alignment.

Primero tendremos que generar un archivo *txt.done.data* donde estén los nombres de cada archivo de audio y su transcripción grafemica. 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[12], Festvox 2,7[13] y speech_tools 2,4[14] 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, copiar 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 simbolos 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 de manera muy cercana a los del castellano río platense (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 te permite crear transcripciones fonéticas 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 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 ($f0$) de utilizando una combinación de HMMs:

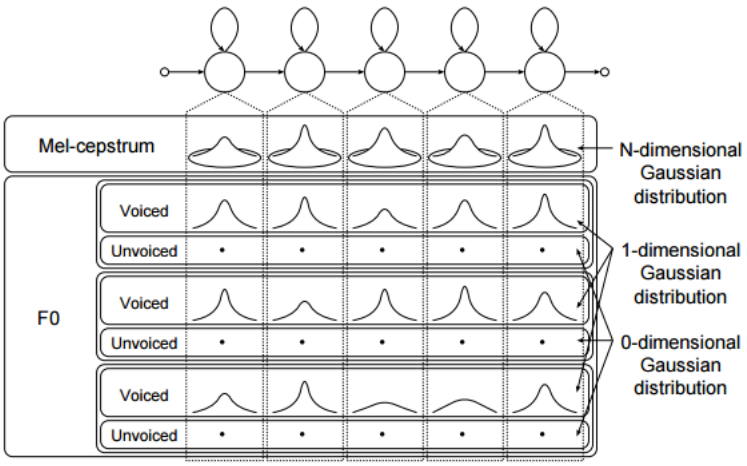


Figure 5.3: structure of HMM.

Fig. 2.1: 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, pag. 41)

Como se ve en la figura 2.1 el espectro y la frecuencia fundamental son modelados en paralelo usando vectores separados. En particular el espectro será modelado como un

vector de gaussianas n dimensional, mientras que la frecuencia principal será modelado como 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 arboles de decisión para clusterizar los datos. Notar que cada hoja del árbol resultante coincidirá con un vector n dimensional de gaussianas o un conjunto de vectores de gaussianas de dimensión cero y uno, según corresponda al espectro o a la frecuencia principal.

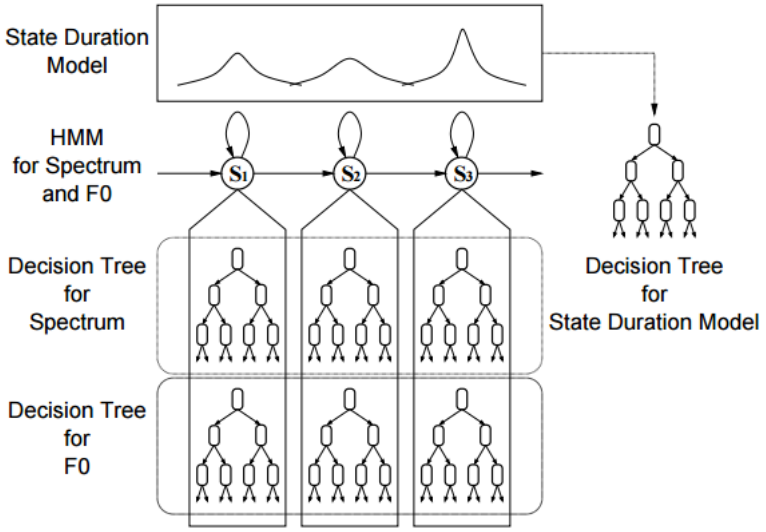


Figure 5.4: Decision trees.

Fig. 2.2: Esquema HMM generado utilizando arboles de decisión (simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura, january 2002, pag. 45)

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

Por otro lado como información contextual para el entrenamiento se tomaron los dos fonemas precedentes y los dos fonemas procedentes para cada fonema y la siguiente información fonética.

- 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 pertenece: por ejemplo para el fonema /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 lennis o fortis.

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

En la imagen 2.3 se muestra el resultado de un fragmento de uno de los arboles de decisión generado para modelar la duración de un fonema. En base a este modelo, el sistema podrá inferir por ejemplo que 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*.

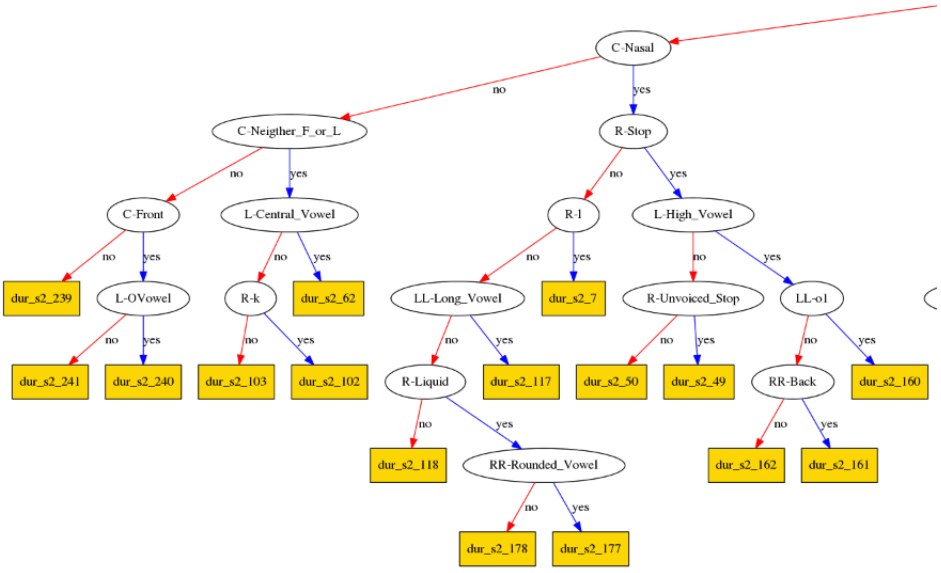


Fig. 2.3: Árbol de decisión generado a partir de los datos para la duración de un HMM

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 mucho peor en los audios generados, sonando estos sumamente metálicos y carentes de prosodia. Esto se debía, posiblemente, a que los arboles 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 con ellas pudimos comprobar que las voces sonaban mucho mas humanas.

2.5.3. Entrenamiento

Desde un punto de vista puramente tecnico, utilizar HTS para entrenar un modelo es bastante sencillo.

Asumiendo que todos los paquetes necesarios fueron instalados, es posible entrenar una nueva voz adaptando la demo disponible en la pagina de descargas de HTS.

Para ello, reemplazar los audios en la carpeta data/raw con aquellos aquellos que se quieran utilizar y los utts correspondientes previamente generados.

Además como adelantamos en la sección anterior, será necesario indicarle a HTS que 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 5.1 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
```

Que extraerá features del audio y construirá los archivos de entrenamiento a partir de los mismos entre otras cosas.

Finalmente para dar comienzo al entrenamiento, ejecutar en la carpeta raíz.

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

Una vez que se complete el entrenamiento, será posible encontrar en la carpeta X el modelo generado (.htsvoice).

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

Además HTS nos pide que explicitemos un rango de extracción para frecuencia fundamental. Tanto para SECYT como para Loc1-Pal el rango 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 permite interpolar con pesos arbitrarios entre varios modelos para producir un nuevo modelo con una mezcla de la carga fonética y prosódica de ambos hablantes y sintetizar audios. Esto nos brinda un gran rango exploratorio y nos permitirá ajustar la carga fonética de los modelos originales para acercarnos al modelo deseado.

A grandes rasgos la interpolación consistirá en tomar los vectores generados anteriormente durante el entrenamiento e interpolar sus funciones de densidad gaussianas para obtener una nueva. En la imagen 2.4 (extraída del trabajo [11]) puede verse la interpolación de N HMMs, cada uno con peso arbitrario a_1, a_2, ..., a_N, que generaran un nuevo modelo Λ.

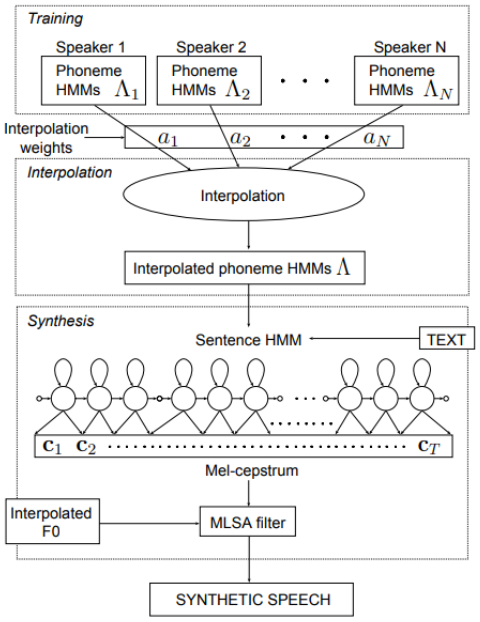


Fig. 2.4: Block diagram of speech synthesis system with speaker interpolaiton.

Una vez obtenido el nuevo HMM es generado el proceso de síntesis puede ocurrir como para cualquier otro modelo, ilustrado en la etapa de síntesis de la figura.

El proceso de síntesis es bastante simple. Asumiendo que todos 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, duración del audio y otros aspectos de la síntesis. La documentación completa puede encontrarse en http://hts-engine.sourceforge.net/.

3. EVALUACIÓN PERCEPTUAL

En este apartado intentamos validar que los modelos generados realmente pueden ser identificados como hablantes extranjeros 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.

Para evitar que el participante pudiera deducir las palabras 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 (presente en el repertorio del castellano) y al menos veinte veces cada fono perteneciente a una vocal.

Los oraciones finalmente 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 uno 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 habrá 5 mezclas diferentes, lo que hace un total de 50 audios sintetizados diferentes.

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 audios distintos.

Nos propusimos como objetivo conseguir 5 respuestas para cada uno de los 50 audios sintetizados, momento en el cual se cierra 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 del estudio. Con la intención de estandarizar los resultados, fue requisito obligatorio utilizar auriculares para la encuesta. También se le pidió a cada participante que la realizara en un lugar silencioso y tranquilo.

3.1. Interfaz

En este apartado 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 principal 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 indique el rango correspondiente a su edad, yendo desde 18 a 25, 26 a 35, y así de diez en diez.

Fig. 3.1: Datos Personales

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.

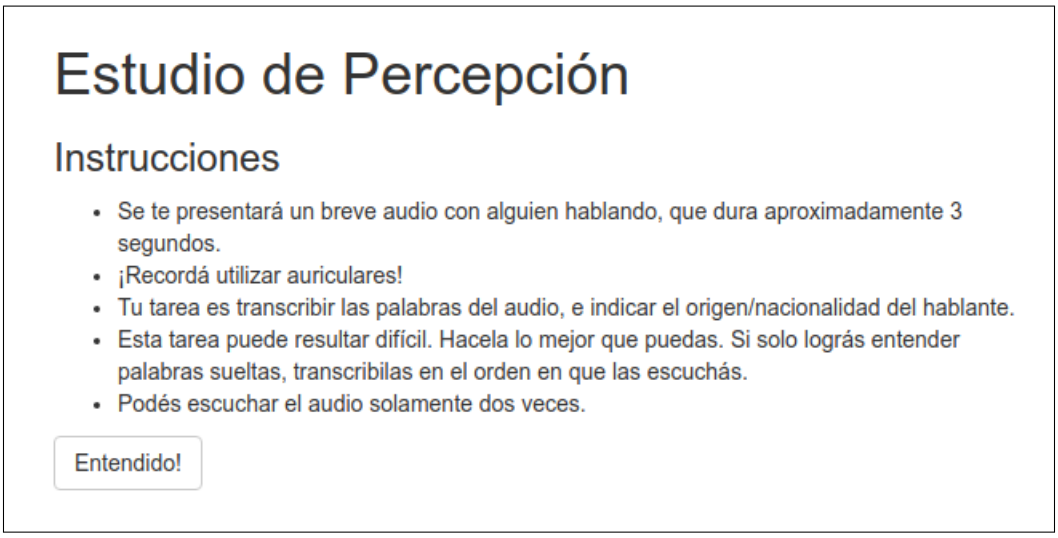


Fig. 3.2: Instrucciones

Como puede verse en la figura 3.2, una vez completados estos datos, a cada participante se le presentó otra vista con las instrucciones especificas 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 el origen de la nacionalidad correspondiente a la voz, como puede apreciarse en la figura 3.3.

Fig. 3.3: 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 (figura 3.4).

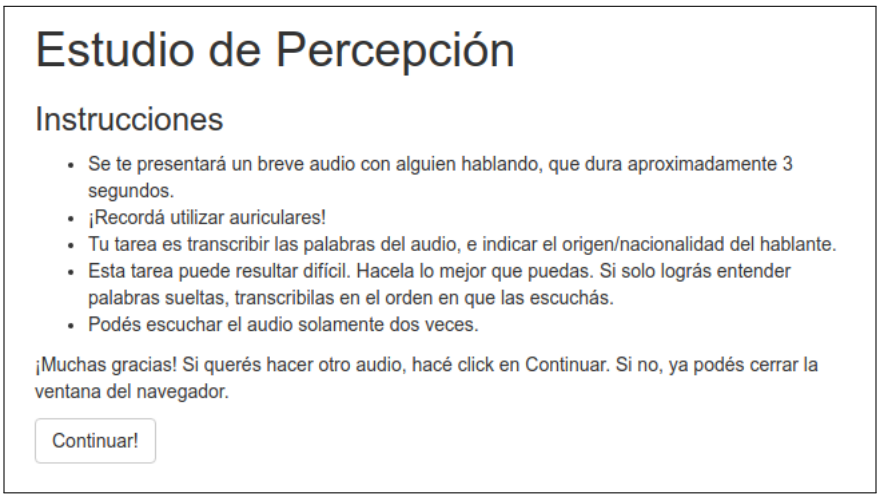


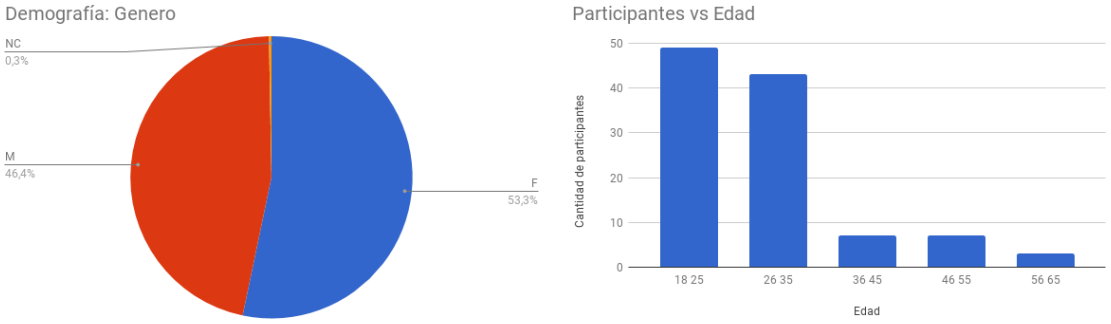
Fig. 3.4: Dialogo Final

3.2. Resultados

A modo de introducción, comenzamos esta sección mostrando los datos demográficos obtenidos. Más adelante, continuamos con un análisis más exhaustivo de inteligibilidad y por separado se realizará otro análisis respecto a la nacionalidad atribuida a las oraciones. Por último para evaluar la hipótesis original, compondremos estos dos ejes para dilucidar el grado de validez de los resultados.

3.3. 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 (fig. 3.5a).



(a) Género *(b)* Edad

Fig. 3.5: Datos demográficos de los participantes

Con respecto al genero de los participantes, 187 respuestas fueron brindadas por participantes del genero femenino mientras que 163 respuestas fueron brindadas por participante del genero masculino (fig. 3.5).

Con respecto de 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 (fig. 3.6).

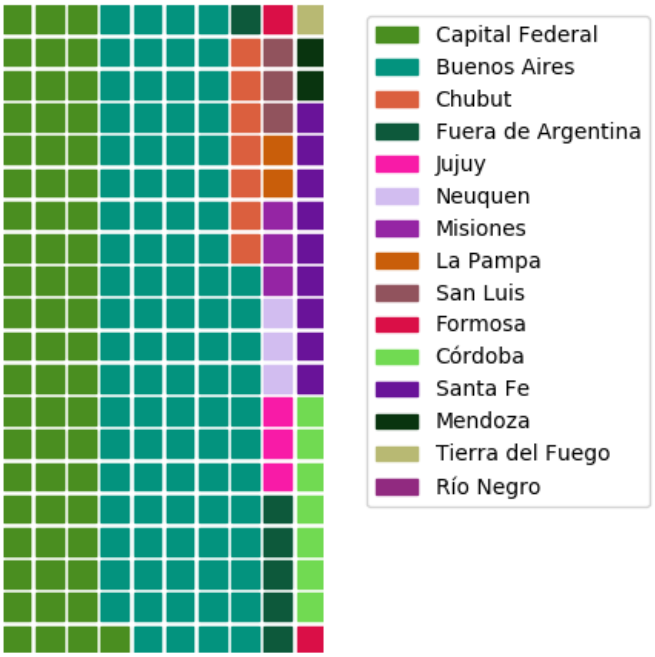


Fig. 3.6: Distribución Territorial

3.4. Inteligibilidad

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

Para esto utilizaremos la distancia de Levenshtein, que consiste en calcular la menor cantidad posible de inserciones, remociones o reemplazos de caracteres que son requeridos para transformar la oración transcripta por un participante en la oración objetivo. 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 3.7 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 irán desde 30 % inglés - 70 % castellano (desde ahora, 30 % ingles) hasta 70 % ingles - 30 % castellano (desde ahora, 70 % ingles). En el eje y presentamos la distancia de Levenshtein entre la oración objetivo y aquella transcripta por 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 (linea interior que atraviesa la caja). Adicionalmente podemos observar con círculos vacíos los outliers de la muestra.

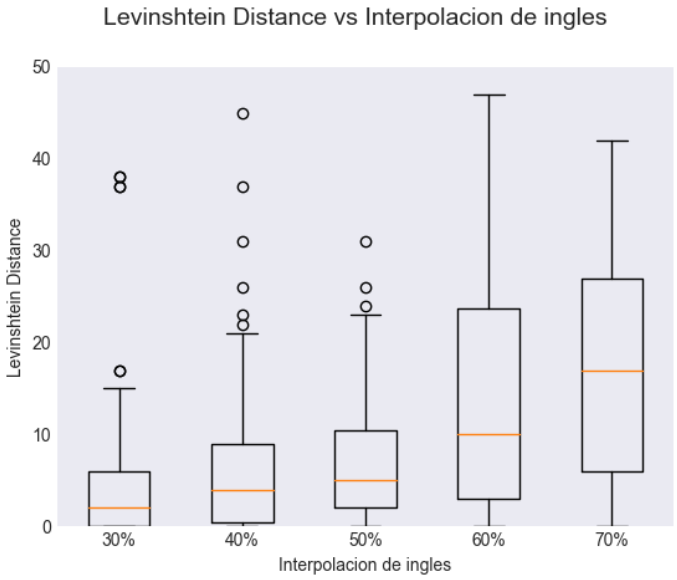


Fig. 3.7: Distancia de Levenshtein para distintos grados de interpolación

Como puede observarse hasta el 50 % de interpolación de inglés, 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 pudimos 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 extremo, 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 y modismos tienen como consecuencia directa que la distancia de Levenshtein se vea afectada. Por ejemplo, un participante que haya escrito "*no entendí nada*" como respuesta devolverá una distancia distinta de aquel que simplemente dejó el campo vacío, cuando en realidad expresan el mismo grado de compresión del texto.

Con el objetivo de reducir esta variabilidad en la muestra, se decidió realizar una

limpieza de los datos. Buscamos uniformizar los datos para un mejor análisis, intentando mantener siempre el espíritu de la respuesta dada por el participante. De esta manera consideramos que si un participante escribió "..." en medio de una oración, lo que quiso decir es que no comprendió parte de la misma. Hubiera sido igual que en su lugar hubiese escrito la cadena vacía "", por ejemplo.

De esta manera 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 alguna ambivalencia, como *concluyo*, no fueron modificadas ya que tanto *concluyó/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.

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

3.6. Datos normalizados

De manera similar que en el apartado anterior, en la figura 3.8 presentamos los resultados esta vez con los datos normalizados.

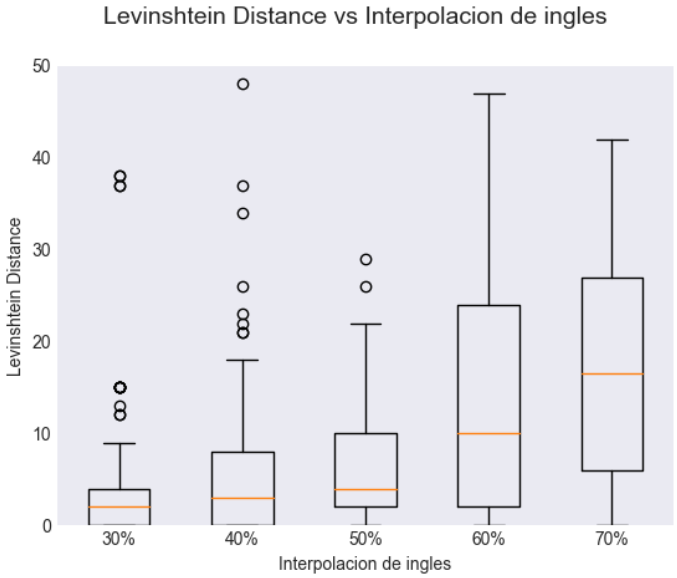


Fig. 3.8: Distancia de Levenshtein para distintos grados de interpolación con datos normalizados

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.

Análisis Estadístico

Con el objetivo de ver que tan estadisticamente significativo son las diferencias en inteligibilidad entre los saltos de interpolación, procedemos ahora a hacer un análisis estadístico sobre los datos.

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

Interpolacion	p-valor	W
30 %	6.875e-13	0.58326
40 %	6.875e-13	0.69505
50 %	1.271e-06	0.86748
60 %	2.932e-06	0.86947
60 %	0.00215	0.93804

Tab. 3.1: Test de Shapiro sobre los datos normalizados

Como puede verse en la imagen 3.1 los 5 tests dan $p < 0{,}01$, entonces rechazamos la hipotesis nula de normalidad, y concluimos que no son normales. Por este motivo descartamos la posibilidad de usar el Student's two-sample t-test y procedemos a utilizar el test no paramétrico Mann-Whitney-Wilcoxon, que solo pide como requerimiento que las observaciones de ambos grupos sean independientes. Cabe recordar que por diseño, nuestras observaciones cumplen este requerimiento, ya que el sistema siempre asignaba audios con diferentes oraciones a un mismo participante, por lo que aunque contestaran mas de una vez, las respuestas resultaban independientes entre si.

Interpolaciones	p-valor	W
30 % vs. 40 %	0.1367	2041.5
40 % vs. 50 %	0.1148	2132
50 % vs. 60 %	0.005142	1921.5
60 % vs. 70 %	0.07083	1956

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

Como se ve en la tabla 3.2 las 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 ingles ($p = 0{,}1148$). En cambio, al pasar de 50 % a 60 % la diferencia es significativa ($p = 0{,}005142$), lo cual significa que en este salto la inteligibilidad se deteriora más sensiblemente que en los saltos anteriores. Por último, de 60 % a 70 % la diferencia de distancias de Levinshtein es aproximadamente significativa ($p < 0{,}1$).

Análisis por oración

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

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 mas comunes varían desde falta de acentos en palabras como “concluyó” hasta faltas de inteligibilidad en palabras con cierta complejidad fonética como “aguileña” o “gruñón”.

Como caso particular la oración 3: “este enjoyado juez comprará nuestro corchete” podemos observar 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 fonemas que realizamos previamente, descubrimos que el mapeo de /hh/ a /g/ era erróneo. Esto produjo que palabras como “gato” sonaran mas como “jato” (/hh/ /a/ /t/ /o/). Adicionalmente descubrimos que ningún fono del inglés estaba siendo mapeado al fono /x/. Esto produjo que cuando el sintetizador interpola entre 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 porque algunos participantes tuvieron problemas transcribiendo la palabra “juez”.

Para los grados de interpolación 60 % y 70 % inglés, podemos observar un aumento notable de la variabilidad en las respuestas. Para el primero, de las 70 respuestas obtenidas, 40 participantes lograron transcribir el audio con una distancia menor a diez caracteres, 6 obtuvieron anotaron una distancia entre 10 y 20 caracteres, y 24 transcribieron el audio con distancia mayor a veinte caracteres. La distancia entre el primer y tercer cuartil pasa a ser de 20 caracteres con una mediana igual a 10.

Para 70 % inglés, 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. La distancia intercuartil se mantiene similar a la de mezcla 60 % inglés, aproximadamente 20 caracteres pero vemos un salto en la mediana que ahora es de 20 caracteres.

Consideramos que este salto en la distancia intercuartil puede deberse a dos motivos: El primero es que existen características particulares de los participantes y sus capacidades para discernir palabras, incluso cuando presentan defectos en la pronunciación del hablante. En particular, la oración 4 (ver figura 3.9) muestra cómo para el 70 % de inglés - 30 % de español en la interpolación, 2 participantes de los 9 que realizaron la transcripción, obtuvieron distancias 2 y 6 en sus transcripciones.

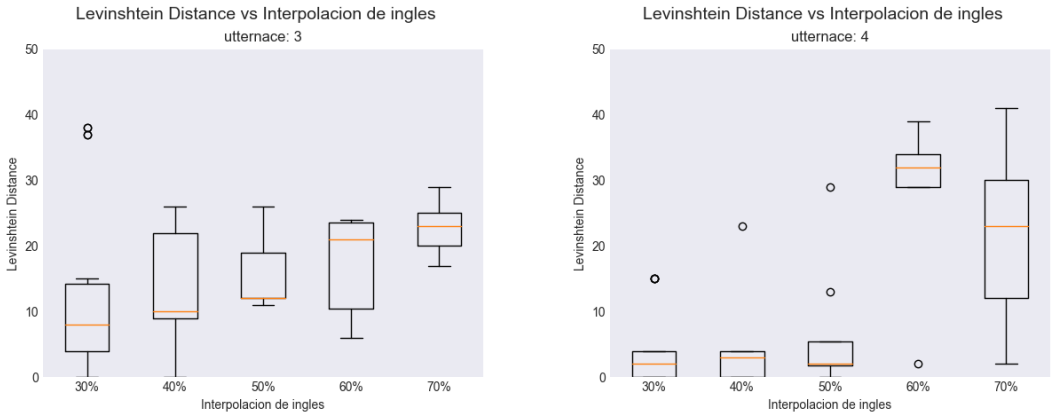


Fig. 3.9: Oración 4 Normalizada

El segundo motivo puede deberse a que existen características particulares de las oraciones o del modelo utilizado para generar la voz que afectan la comprensión del audio: la oración 10 "El nudillo Argentino perdió su vaso" con 70 % de interpolación de inglés - 30 % de castellano, en la cual 6 de los 8 participantes obtuvieron una buena transcripción del audio con distancia menor a 15 [1], y la oración 8 "Las acongojadas cotorras sonrieron a mi círculo", donde, para ese mismo grado de interpolación, todos los participantes transcribieron el audio con distancia de Levenshtein mayor a 20, parecen demostrar esto. O bien la dificultad de las oraciones 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.

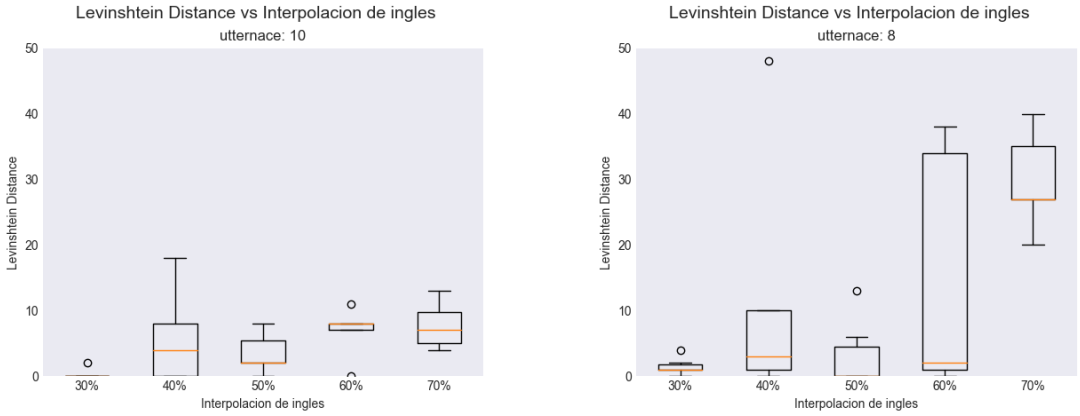


Fig. 3.10: Oración 10 y 8 Normalizados

En conclusión, en esta sección pudimos demostrar que fue posible generar una voz con una distancia menor a 10 caracteres hasta un 50 % de inglés y 50 % de castellano. Pasado el 50 % de inglés, la variabilidad de las respuestas se vuelve mucho mas grande pudiendo

[1] Incluso teniendo en cuenta que esta oración se ve afectada por el mapeo incorrecto de la /g/ como fue discutido previamente

haber participantes que anotan una buena distancia de Levenshtein (10 caracteres o menos) hasta algunos que no logran comprender ni siquiera segmentos aislados del mismo (mas de 30 caracteres).

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 la voz.

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:

- Hispanoparlate: "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 3.11 presentamos las nacionalidades atribuidas a la voz generada para cada punto de la interpolación.

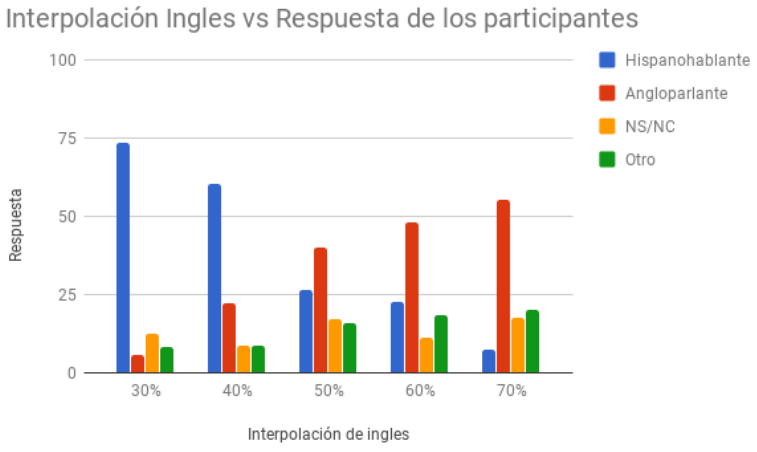


Fig. 3.11: Análisis General

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 Hispanohablante. Dicho de otra manera, los datos parecen sugerir que con cada salto en la interpolación inglés, la cantidad

de participantes que afirma que la voz pertenece a un Hispanohablante, disminuye en aproximadamente 7 puntos conceptuales.

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 ingles.

Análisis Estadístico

Para darle peso estadístico a las afirmaciones realizadas en el aparatado 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 ingles. 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 interpolacion de inglés en el modelo y como variable dependiente tenderemos el porcentaje de sujetos que percibieron el audio como un hablante español.

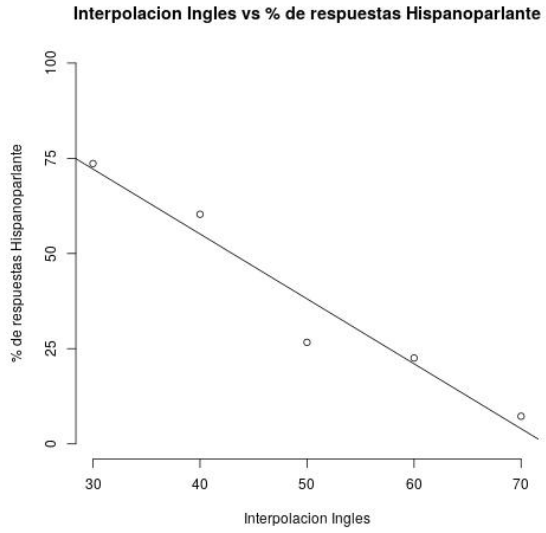


Fig. 3.12: Modelo lineal ajustado para la percepción de hispanoparlante

Como puede verse en la figura 3.12 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 interpolacion de inglés y como variable dependiente tendremos el porcentaje de sujetos que percibieron el audio como un hablante inglés.

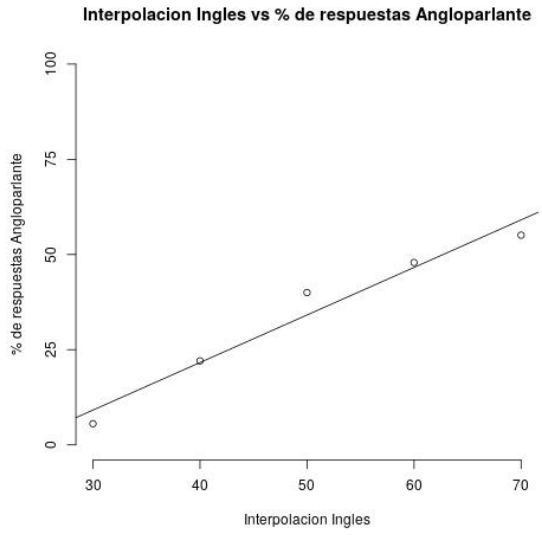


Fig. 3.13: Modelo lineal ajustado para la percepción de angloparlante

En la figura 3.13 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 estadistica que indica que aumentar el grado de interpolación resulta en mas participantes reconociendolo como una persona de habla inglesa y no como un hablante oriundo de un país de habla hispana.

Análisis por oración

Observando las oraciones una por una, podemos 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 ingles, el 60,00 % de los participantes considera que la voz pertenece a un hablante de habla anglosajona, siendo que en los resultado generales, para ese mismo grado de interpolación, menos del 25 % de los participantes lo afirmaban.

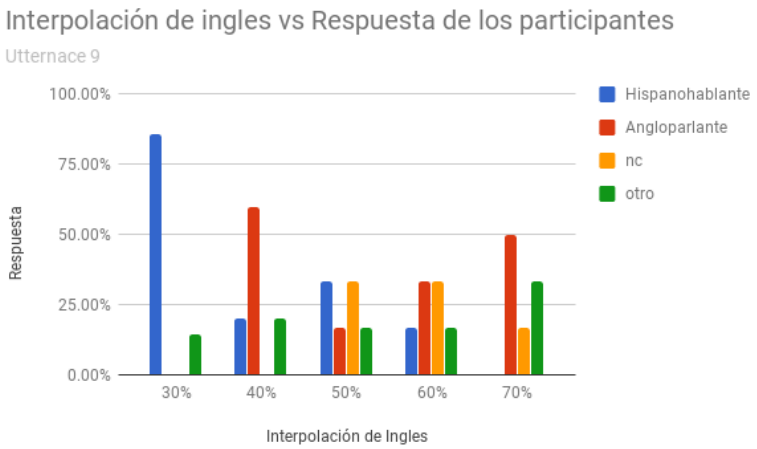


Fig. 3.14: Oración 3 y 9

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 un /r/ (*perro*) que resulta muy notorio al pronunciarse con una intensidad menor a la esperada (mas similar a una /ɾ/ (*pero*)) y puede ser atribuida, entre otras causas, a un hablante Anglosajón con dificultades en la dicción de fonemas extranjeros.

Bajo esta suposición buscamos en que otras oraciones se presenta este fonema:

- 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 3.15 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.

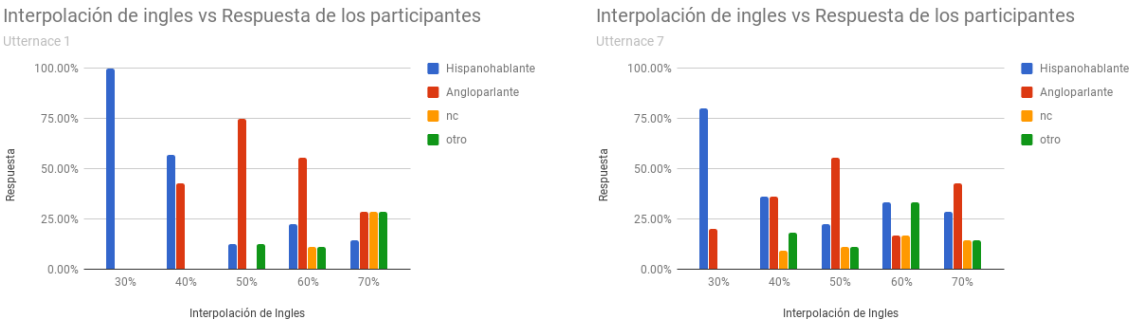


Fig. 3.15: Oración 1 y 6

Por otro lado en la figura 3.16 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 mas difusa.

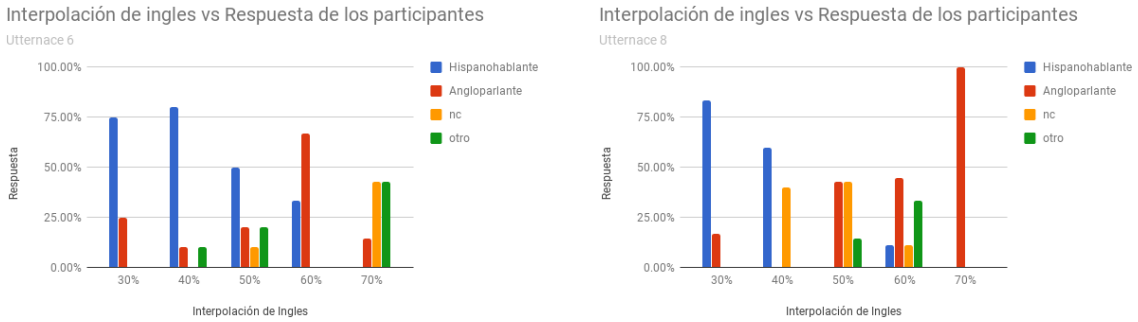


Fig. 3.16: Oración 7 y 8

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.

3.8. Resultados Generales de la experimentación

En la figura 3.17 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 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 distancia media (linea interior que atraviesa la caja). 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. Adicionalmente podemos observar con círculos vacíos los outliers de la muestra.

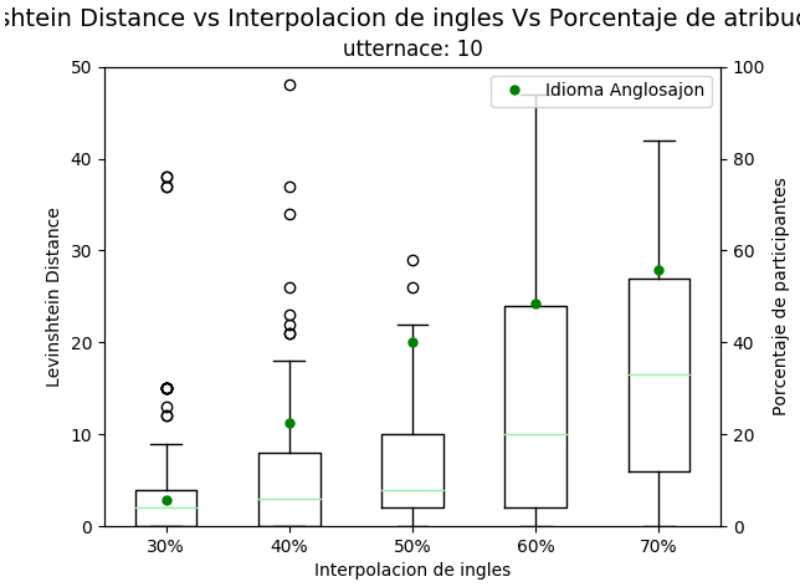


Fig. 3.17: Distancia de Levenshtein vs distintos grados de interpolación vs Porcentaje de participantes que consideran hablante anglosajón

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 valido 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. CONCLUCIONES Y TRABAJO FUTURO

Como trabajo futuro queda corregir los problemas en el mapeo fonético encontrados en la fase de experimentación.

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 problemáticos como el caso de la /r/ vibrante el grado de interpolación podría dejarse más cercano al castellano, mientras que para fonemas con comportamientos más similares el grado de interpolación podría llevarse más cerca del modelo ingles.

5. APENDICE

5.1. questions_qst001.hed

Bibliografía

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

[2] Automatic determination of phrase breaks for Argentine Spanish, Humberto M. Torres & Jorge A. Gurlekian, Laboratorio de Investigaciones Sensoriales CONICET, University of Buenos Aires, Argentina

[3] Torres, Humberto & Gurlekian, Jorge & Cossio-Mercado, Christian. (2012). Aromo: Argentine Spanish TTS System.

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

[5] 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

[6] Simultaneous Modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura, pag. 28

[7] Speaker Similarity Evaluation Of Foreign-accented Speech Synthesis Using Hmm-based Speaker Adaptation, Mirjam Wester, Reima Karhila

[8] Simultaneous Modeling of phonetic and prosodic parameters, and characteristic conversion for hmm-based text-to-speech systems, takayoshi yoshimura

[9] 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/ICASSP2006/0100853.pdf

[10] Transcription of Spanish and Spanish-Influenced English, Brian Goldstein

[11] Speaker Interpolation in HMM-based Speech Synthesis System, Takayoshi Yoshimura1, Takashi Masuko2, Keiichi Tokuda1, Takao Kobayashi2 and Tadashi Kitamura1

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

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

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