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

Implementación y evaluación de un sistema de sintesis de habla con acento extranjero variable

Tesis de Licenciatura en Ciencias de la Computación

Franco Negri

Director: Agustin Gravano
Codirector: Master Yoda
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 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. Si bien actualmente se considera que el estado del arte para la síntesis es el entrenamiento con redes neuronales profundas (DNN), una técnica todavía utilizada es la que utiliza 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.

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, algo que puede volverse mas complicado cuando se esta trabajando con redes neuronales profundas.
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. 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, pasando por cuestiones lingüísticas, como es poder vislumbrar el limite en que un acento deja de parecernos local para pasar a ser extranjero y cuestiones psicológicas: lograr distintos efectos sobre el usuario, quien podría reaccionar distinto ante diferentes acentos.
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 disimiles (para este caso de estudio: castellano e ingles)
Para este trabajo nos basaremos fuertemente en la síntesis/análisis mel-cepstral, 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 Profesor Tadashi Kitamura, Nagoya Institute of Technology[7].

Como introducción a este trabajo comenzaremos analizando las decisiones de metodología utilizadas a lo largo de la investigación, así también como detalles teóricos como el mapeo de fonemas necesario para adaptar el repertorio fonético del ingles al castellano, etc.

En las siguiente secciones se detallaran distintos temas que fueron necesarios abordar para llevar a cabo esta tesis. En orden de aparición estos son:

1. Realizar un etiquetado fonético de distintos corpus de audios.

2. Realizar un mapeo entre los fonos del castellano y los del ingles. Estos serán necesarios en el paso siguiente donde será requerimiento indispensable tener modelados

los mismos fonos para todos los modelos.

3. Realizar el entrenamiento de los HMM+GMM. Para esto contaremos con el framework de modelado de HMMs HTS.

4. Utilizar las herramientas provistas por HTS para interpolar entre modelos y poder sintetizar habla con distintos grados de fonética y prosodia inglesa.

Para finalizar el apartado teórico discutiremos algunos aspectos implémentatelos de HTS y las otras herramientas utilizadas en este trabajo.

2. MATERIALES Y MÉTODOS

2.1. Preparación De los datos

2.1.1. Primer Corpus En Castellano

Al comienzo de la investigación comenzamos con tan solo un corpus de datos, *secyt-mujer*[1]. Este esta compuesto por 741 oraciones cortas declarativas habladas por una locutora profesional de origen Español Rioplatense, equivalentes a 48 minutos de habla.

A continuacion 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. Estas consistirán principalmente de una lista de fonos donde se indica donde comienza y donde termina cada uno. La calidad de las oraciones que logremos sintetizar a posteriori dependerá fuertemente del etiquetado, por lo que es necesario prestar especial cuidado en que estas estén lo mejor alineadamente posible con los audios, ya que estas son necesarias para entrenar los modelos de HMM+GMM, extrayendo de aquí tanto la información acústica para cada fono (cosas tales como al frecuencia principal, la duración, etc) como así también información contextual (como por ejemplo: como suena un fonema cuando está seguido de algún otro, al principio de una oración, si se encuentra en un diptongo). Por esto una mala transcripción se traducirá indefectiblemente en un mal modelo y malas oraciones sintetizadas.

Realizamos varias pruebas de concepto utilizando HTS y este corpus, experimentando con diversos métodos para obtener el etiquetado fonético.

La primera estrategia consistió en utilizar alineamiento automático utilizando EHMM alignment [8] utilizando Festival y Festvox. Estos programas tienen como ventaja que tanto la anotación del corpus como la generación de trazas fonéticas (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.

Aún así, los resultados preliminares con este método fueron bastante adversos: los audios generados resultaban poco inteligibles notandose claros defectos acústicos, El más notable siendo el fono /r/ que se asemejaba mas a /R/.

Utilizando Praat para visualizar el alineamiento entre utternaces y audios, descubrimos que la alineación estaba desfasada algunas milésimas de segundo. Sospechamos que esto se debió a algún problema con la normalización de los audios.

Dado que para el corpus contamos con las transcripciones fonéticas anotadas de manera manual, procedemos a implementar un híbrido con EHMM. En este híbrido tomamos las anotaciones hechas a mano para cada fonema y la información contextual y el repertorio fonético generado a partir del proceso de EHMM. 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 estos utternaces mixtos resulto ser superior a los generados solo con alineamiento automático. Aún así los audios sintetizados todavía no alcanzan una

calidad aceptable, realizando pruebas internas 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 fraces declarativas y interrogativas del 80 % y 20 % aproximadamente, pronunciadas por una locutora profecional con ascento 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.

Para este corpus no cotábamos 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*. 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 parámetros que consideramos podían afectar la calidad de la voz, logramos obtener resultados que superaban de manera significativa aquellos obtenidos previamente con el otro corpus. 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 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 Ingles

Por otro lado entrenamos una voz en ingles *CMU-ARCTIC-SLT*[3] con 1132 oraciones en ingles y 56 minutos de habla, disponible en la pagina de hts [9]. 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.

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

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

Una lista extensiva de los parámetros utilizados para el entrenamiento se puede ver en el apéndice 3.

2.2. Repertorio Fonetico y Mapeo De Fonemas

Para la generación de utternaces, tanto de los audios en ingles como en castellano, utilizamos los repertorios fonéticos brindados por festvox (ver apéndice 1).

El primer desafío que se presenta es que estos repertorios fonéticos no tienen un mapeo directo con el Alfabeto Fonético Internacional: por ejemplo con este repertorio fonético, en el castellano existen tres fonos distintos para la /i/. Esta decisión 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: /ia/, /ie/, /io/, /iu/.

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

Tab. 2.1: Repertorios Foneticos utilizados por Festvox para Ingles y Castellano

De aquí surgen varios problemas, el primero de los cuales es que festival utiliza el mismo simbolo para dos fonos distintos. De esta manera, la /g/ en castellano es suena mucho mas disminuida que la /g/ del repertorio ingles, aunque ambas sean descritas como un fono oclusivo velar sonoro.

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

Ingles	Castellano
ae	a
aa	a1
b	b
ch	ch
d	d
dh	d
eh	e
el	e1
f	f
hh	g
iy	i
ih	i1
k	k
l	l
jh	ll
m	m
n	n
nx	n
ao	o
ou	o1
p	p
r	r/rr

Ingles	Castellano
s	s
t	t
uw	u
w	u0
uh	u1
dx	-
em	-
en	-
er	-
ei	-
g	-
hv	-
ng	-
th	-
v	-
y	-
sh	-
zh	-
z	-

Tab. 2.2: Mapeo Fonetico

Por otro lado para varios fonos tuvimos que hacer reglas especiales ya que no contabamos con ningun fono del ingles lo suficientemente similar. Así, para el fono ny (ñ o en ipa) colapsamos las apariciones del fono /n/ seguido de /j/. 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*:

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., /stɑ:z/ *stars* → [tɑ:z]) or *epenthesis* (or *prothesis*) (e.g., /stɑ:z/ *stars* → [estɑ:z]) (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, realiza una aproximación entre los fonos conocidos y los fonos 'objetivo' de la nueva lengua.

De manera similar el ingles carece del fono vibrante múltiple alveolar sordo /r/ y dado que el fono /R/ ya estaba siendo utilizado, no podíamos realizar un mapeo tan

directo. Como sulución tomamos la mitad de los fonos /R/ y los remplazamos con /r/. Un problema similar surge del fono /g/, que si bien este existe en ingles, al utilizarlo para sintetizar oraciones sonaba muy alienigena. Con el objetivo de suavizar el sonido tomamos el fono de la /g/ y lo mapeamos a un fono que no fuera utilizado y tomamos los fonos etiquetados con /hh/ del ingles y los remplazamos con el /g/ del castellano.

Aquellos fonos que consideramos suficientemente disimiles 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 sintesis.

Utilizando estos fonos a la hora de entrenar nos permite sintetizar oraciones en castellano, aunque como es de esperar, dado que el corpus de entrenamiento es tan disimilar de las oraciones que queremos sintetizar, donde tanto las combinaciones de fonemas, las reglas prosodicas y las acentuaciónes vocalicas son distintas, los audios sitetizados resultan incomprensibles y de muy baja calidad

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.

Para grados cercanos al 90 % de castellano - 10 de ingles obtenemos los resultados esperables: las oraciones sintetizadas tienen un marcado acento castellano. Así mismo, en el otro extremo, 10 % de castellano - 90 % de ingles, la vos sintetizada, al igual que lo que se describió en el apartado anterior, la voz 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 ingles y 40 % de castellano - 60 % de ingles, podemos apreciar resultados mas cercanos a los esperados, pudiendo apreciar en las oraciones sintetizadas detalles distintivos como el fono /R/ mas suavizada, o pronunciación de vocales mas abiertas, pero aun 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 ingles 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 técnica permite tomar un modelo ya entrenado y adaptarlo para asimilar características de un nuevo hablante. Esta técnica nace de la idea 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. Las pruebas no son concluyentes ya que las oraciones sintetizadas no solamente pierden la identidad del hablante original sino también sus características fonéticas. Si bien existen indicios que indican que es posible generar un modelo con las características deseadas, 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 experimentación.

2.5. Herramientas

En esta sección presentamos las herramientas utilizadas para este trabajo.

2.5.1. Festival y Festvox

Festival es un framework que permite sintetizar habla. Además posee 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 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 Utternaces 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.

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

Como veremos mas adelante, estos utternaces serán utilizados en el entrenamiento con HTS para modelar cada uno de los fonemas.

2.5.2. HTS

HTS es un framework de entrenamiento y sintesis 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:

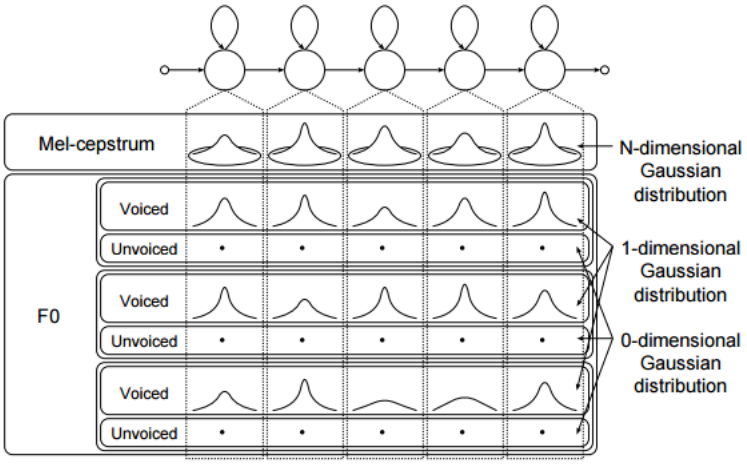


Figure 5.3: structure of HMM.

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)

Mel-celpstum (espectro) y los tres vectores del $f0$ (frecuencia principal) son modelados en paralelo.

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 extraída de los audios de entrenamiento. A continuación se presenta una vista esquemática de la estructura del HMM generado clusterizado con arboles de decisión:

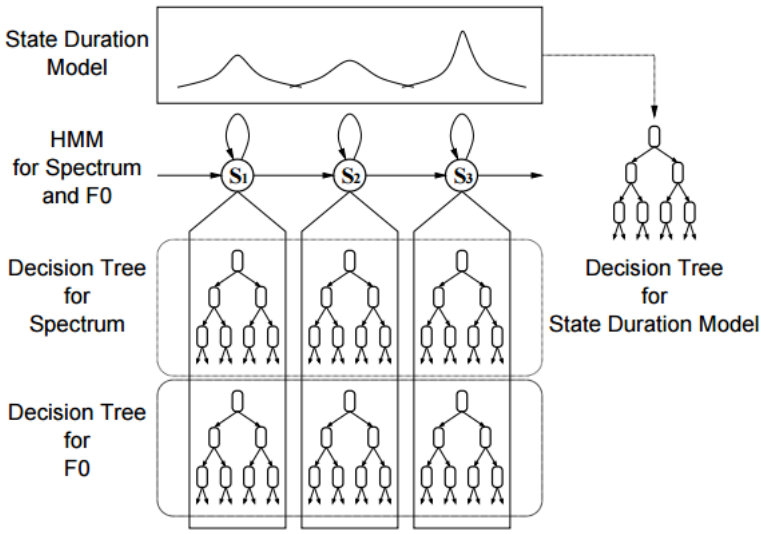


Figure 5.4: Decision trees.

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)

En particular para este trabajo la clusterización de datos se realizó generando arboles de decisión y para cada fonema se tomó los dos fonemas precedentes y los dos fonemas procedentes y se extrajo la siguiente información contextual.

- 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.

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

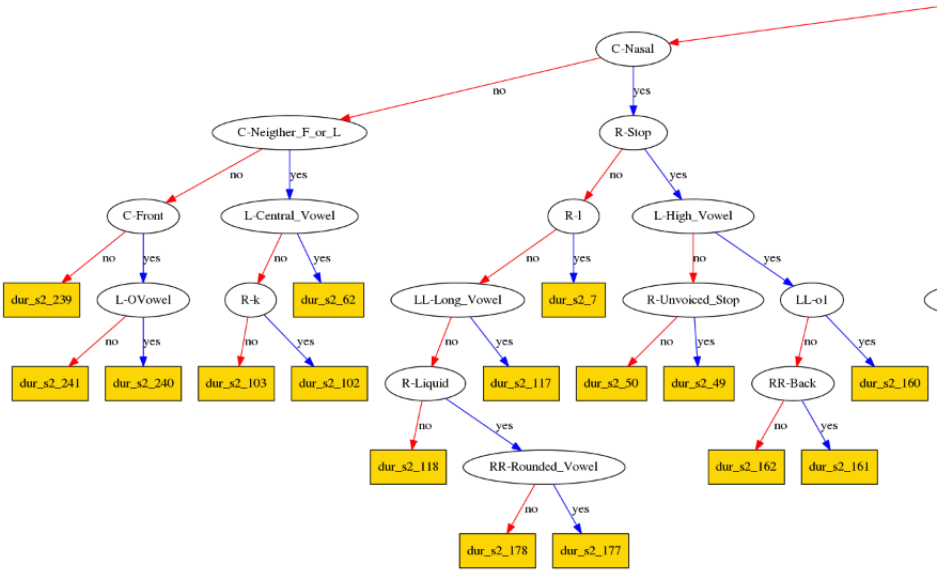


Fig. 2.3: Árbol de decisión generado para la duración de un HMM

Con este modelo, el sistema podrá 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 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. 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. 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.

3. EVALUACIÓN PERCEPTUAL

En la siguiente aparatado intentamos 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 a cada participante se le presentó una oración sintetizada con distintos grados de mezcla de español e ingles 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 son generadas de manera semánticamente impredecible. Esto significa que a partir de una lista de sustantivos, adjetivos, determinantes y verbos se generan oraciones de manera aleatoria con la estructura:

Determinante Adjetivo Sustantivo Verbo Determinante Sustantivo

Luego, para asegurarnos de estar cubriendo todos los posibles fonos del castellano, las oraciones son modificadas para ser fonéticamente balanceadas. Esto querrá decir 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 estos diez oraciones se varió el nivel de mezcla entre 30 % de ingles, 70 % castellano hasta 70 % de ingles, 30 % de castellano, 10 % cada vez. De esta manera, para cada oración habrá 5 mezclas diferentes, lo que hace un total de 50 audios sintetizados diferentes.

La encuesta se realiza 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).

El objetivo de la misma es conseguir para cada uno de los 50 audios sintetizados, 5 respuestas, momento en el cual se cierra la posibilidad de contestar.

La misma se lleva 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.

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 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 pagina principal en la que todos los participantes fueron recibidos.

Fig. 3.1: Datos Personales

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.

Se les pidió, además, que indicara su genero: masculino, femenino, otro, no contesta

y la provincia donde pasó sus primeros diez años de vida. Consideramos que estos datos son importantes para el estudio ya que dependiendo de ellos los resultados variarán indefectiblemente, 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.

Una vez que completados estos datos, se le presenta otra vista con las instrucciones especificas para completar la encuesta (figura 3.2).

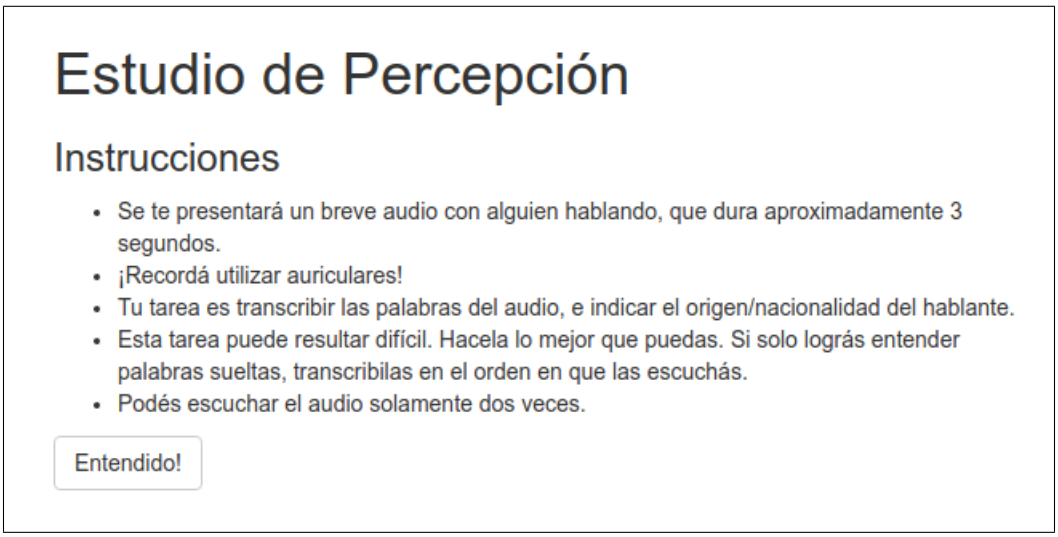


Fig. 3.2: Instrucciones

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 (figura 3.3)

Fig. 3.3: Transcripción

Una vez que la respuesta es guardada, se le preguntaba si quiere continuar transcribiendo otro audio, o de caso de haber contestado cinco veces, se le presentaba un mensaje donde se le indicaba que ya podía cerrar la encuesta (figura 3.4).

Fig. 3.4: Dialogo Final

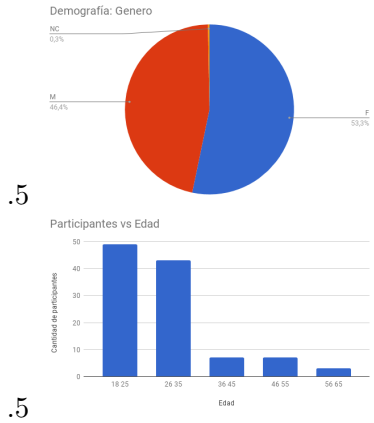
3.2. Resultados

A modo de introducción, comenzaremos esta sección mostrando los datos demográficos obtenidos. Mas adelante, continuaremos con un análisis mas 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 su grado de validez.

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. **??**).



.5 .5

Fig. 3.7: A figure with two subfigures

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 (fig. **??**).

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 (fig. 3.8).

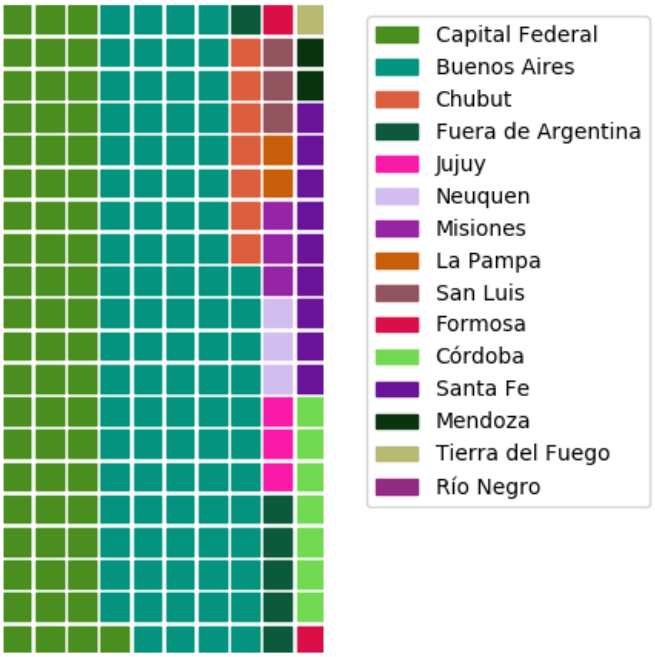


Fig. 3.8: Distribución Territorial

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 cuan 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 posibles 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*, reemplazar *s* por *l* y remover la *a* por lo que la distancia de Levinshtein entre estas dos oraciones es de 3. Además, se considerará un remplazo 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 3.9 presentamos los resultados generales obtenidos sin ningún tipo de modificación.

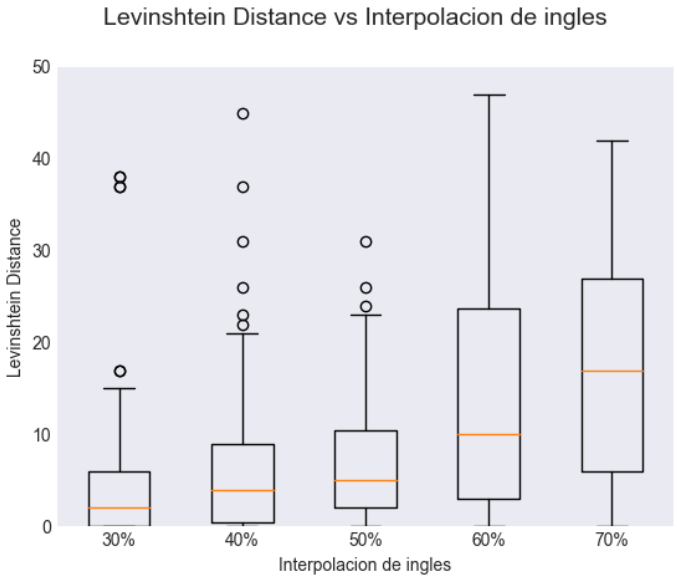


Fig. 3.9: Resultados Generales

Como puede observarse hasta el 50 % de mezcla castellano-ingles, la distancia entre el primer y tercer cuartil menor a 10 caracteres, siendo la media de 5 caracteres. Pasados el 60 % de ingles, se observa un aumento brusco el la distancia intercuartiles, la distancia entre el primer y tercer cuartil pasa a ser cercana a 20 caracteres, y la media 10 caracteres en el caso de 60 % ingles y 15 en el caso del 70 %. En las proximas 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 el 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".

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: comas, puntos, "(?)"
- Reemplazo de oraciones como *blabla*, *no entendí* o cualquier otra expresión que indique ininteligibilidad de una palabra u oración por "".
- Corrección de acentos en palabras no ambiguas: *botón*, *prefirió*, *recorrió*, *chupetín*, *riñon*, *gruñón*.

Aquellas palabras que presentan ambivalencia, como : *concluyó* no fueron modificadas ya que tanto *concluyó/concluyo* son validas. El participante podría haber interpretado la palabra con cualquiera de las dos connotaciones cambiando el significado de la interpretación.

Esperamos que esta normalización no solamente ayude a disminuir la propagación de los resultados, sino que además, nos permitirá interpretar de manera mas intuitiva el significado la distancia de Levenshtein en cada caso.

3.6. Datos normalizados

Podemos observar en la figura 3.10 los nuevos resultados con los datos normalizados. Para 40 % y 50 % ingles . Para 60 % . Para 70 % .

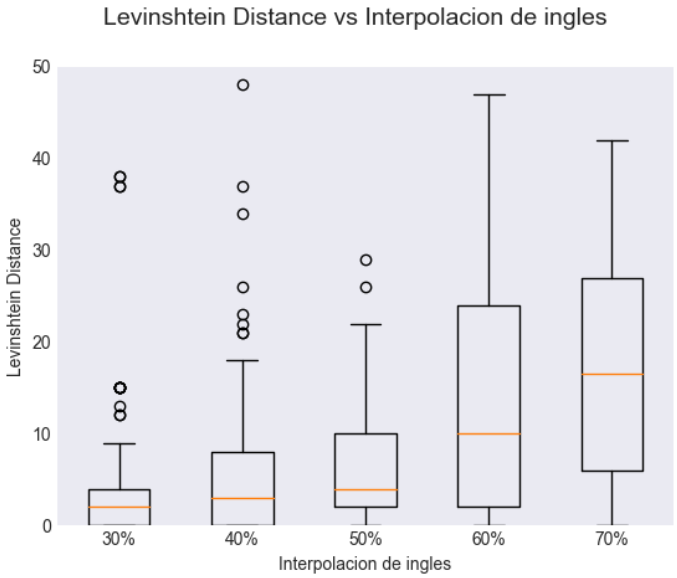


Fig. 3.10: General Normalizado

Con este baseline, podemos ver que para una interpolación de ingles 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 media cercana a 2

Para la mezcla 40 % ingles - 60 % castellano, de un total de 67 participantes, 57 anotaron una distancia de Levinshtein 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 media muy similar a la mezcla 30 % ingles.

Para la interpolación 50 % ingles - 50 % castellano, 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 media es de 3 caracteres.

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”, y “enjoyado” que transcribieron como “enfollado”, “enrollado” y la conjugación exacta del verbo comprar.

Para los grados de interpolación 60 % ingles - 40 % castellano y 70 % ingles - 30 % castellano, 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 media igual a 10.

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. La distancia intercuartil se mantiene similar a la de mezcla 60 % ingles, aproximadamente 20 caracteres pero vemos un salto en la media 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 3.11) muestra como para el mismo grado de interpolación, con características similares 2 participantes de los 9 que realizaron la transcripción, obtuvieron distancias 2 y 6 en sus transcripciones.

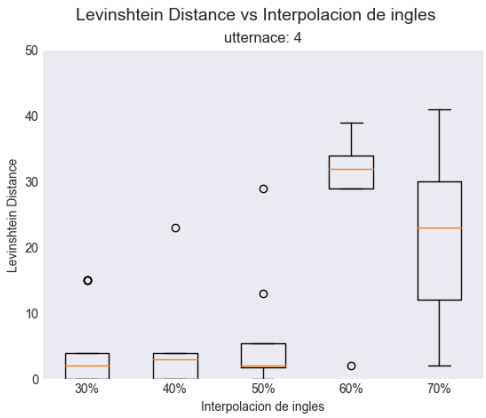


Fig. 3.11: Oraciones 3 y 4 Normalizados

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, donde el 6 de los 8 participantes obtuvieron una buena transcripción del audio, y la oración 8, donde todos los participantes transcribieron el audio con inteligibilidad baja o nula, 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.

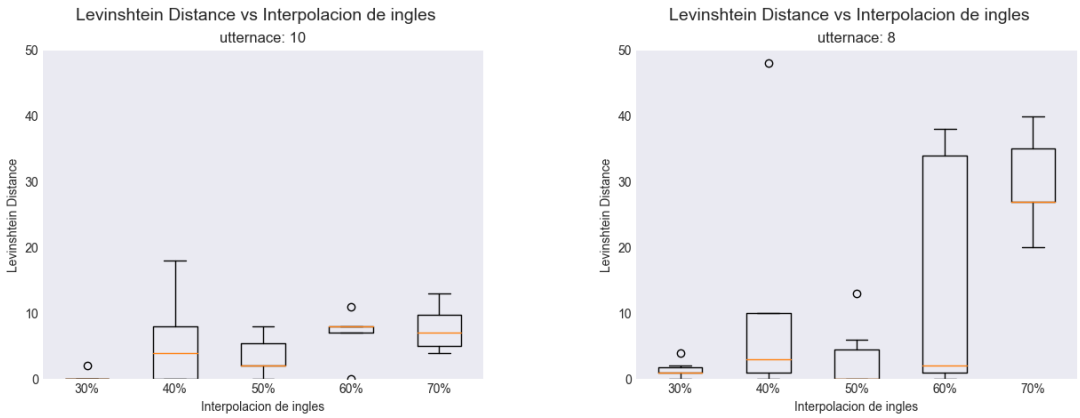


Fig. 3.12: Oración 10 y 8 Normalizados

En conclusión, en esta sección pudimos demostrar que fue posible generar una voz con una distancia menor a 10 caracteres hasta un 50 % de ingles y 50 % de castellano. Pasado el 50 % de ingles, la variabilidad de las respuestas se vuelve mucho mas grande pudiendo haber participantes que anotan una buena distancia de Levinshtein (10 caracteres o menos) hasta algunos que no logran comprender ni siquiera segmentos aislados del mismo (mas de 30 caracteres).

3.7. Análisis del Órigen Percibido

En esta sección analizaremos los resultados de las 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 tomaron la consigna de manera diferente, 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 son validas ya que no aportan información para esta investigación.

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

- 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, en la figura 3.13 presentamos las nacionalidades atribuidas a la voz generada para cada punto de la interpolación.

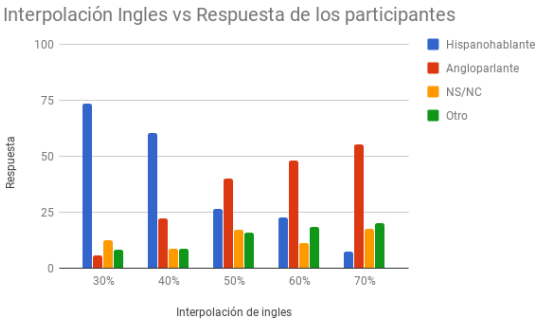


Fig. 3.13: Análisis General

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.

Con 50 % y 60 % de ingles los resultados son similares. obtenemos que aproximadamente en el 50 % de las oraciones, 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 las oraciones 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 las oraciones se puede apreciar que al menos 50 % de los participantes dijo que el hablante era de origen anglosajón. Mas aún, en el 40 % de las oraciones 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 las oraciones ningún participante considera que la voz sea de habla hispana. En el 30 % restante, 25 % de los participantes o menos consideran que la voz pertenezca a un hispanohablante.

Observando las oraciones una por una, podemos observar algunas particularidades. para 40 % ingles, puede verse que no hay una votación homogénea. Por ejemplo en la figura 3.14 en la oración 3 el 80 % de los participantes coincide que la voz pertenece a un hablante de habla hispana, mientras que en la oración 9 el 60,00 % de los participantes considera que la voz pertenece a un hablante de habla anglosajona.

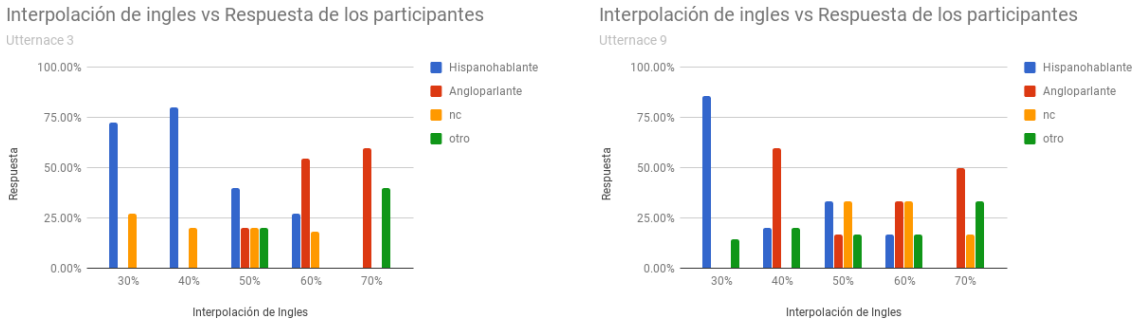


Fig. 3.14: Oración 3 y 9

Esta gran disparidad de resultados entre distintas oraciones se puede atribuir a las características particulares de cada oración. En particular la oración 9: “Ese gruñón perro prometió a esos cuñados” contiene una /r/ que resulta muy notoria al pronunciarse con una intensidad menor a la esperada (mas similar a una /R/) y es atribuida, en general, a un hablante extranjero.

Bajo esta suposición observamos que las otras oraciones que presentan 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”

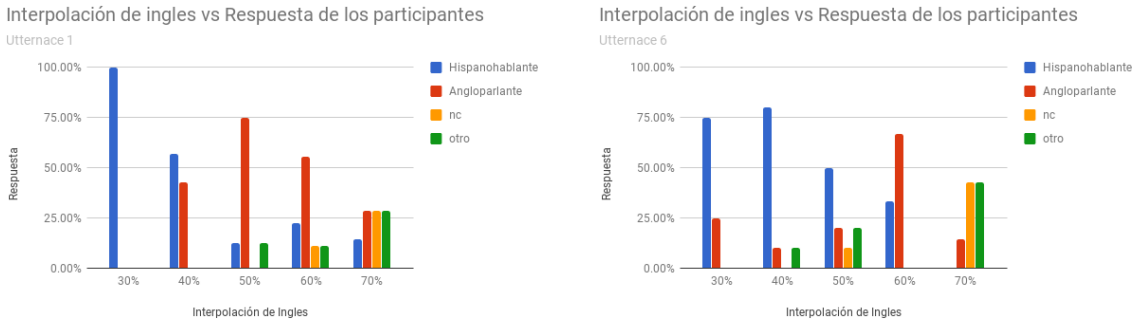


Fig. 3.15: Oración 1 y 6

También presentan un mayor porcentaje de atribuciones a nacionalidad anglosajona.

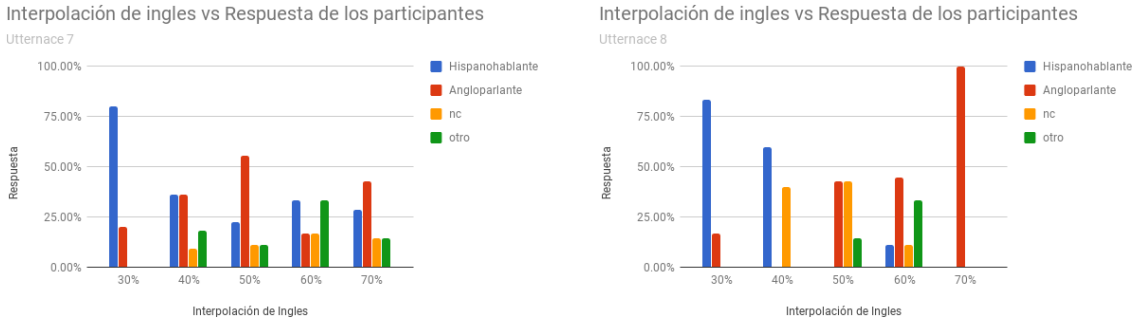


Fig. 3.16: Oración 7 y 8

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

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

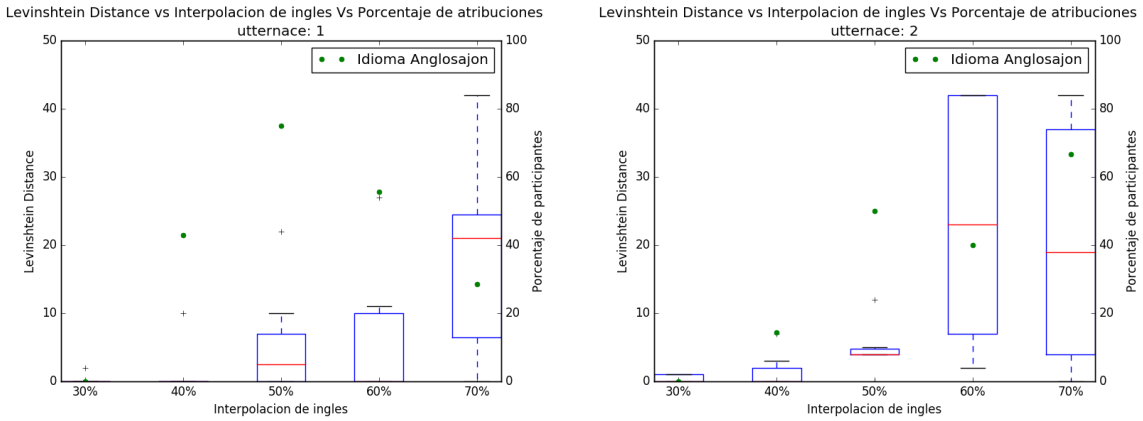


Fig. 3.17: Oración 1 y 2

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 de la oración elegido y el grado de interpolación.

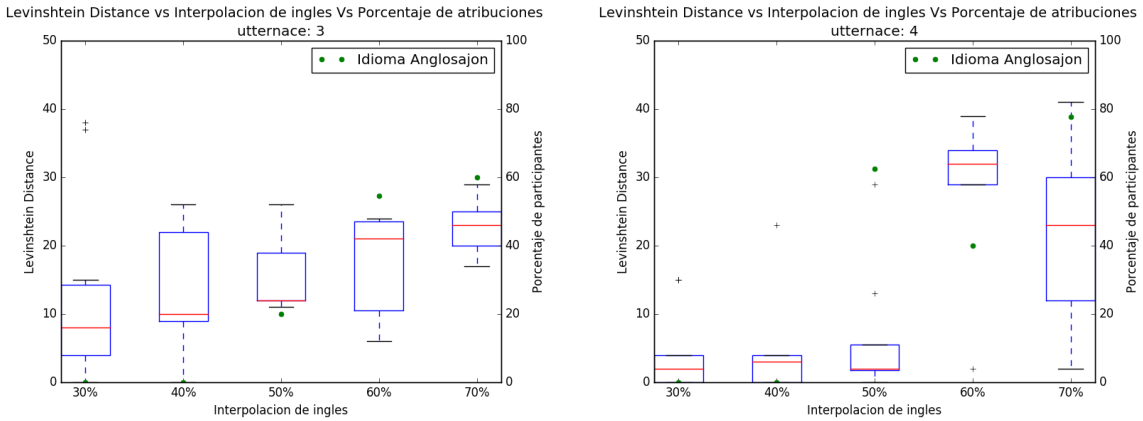


Fig. 3.18: Oración 3 y 4

También es interesante observar casos como el que se presenta comparando el oración 10 y la oración 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 sonoridad de ciertos fonemas o la prosodia general de la voz.

Este no es un caso aislado, véase que lo mismo sucede con la oración 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 %.

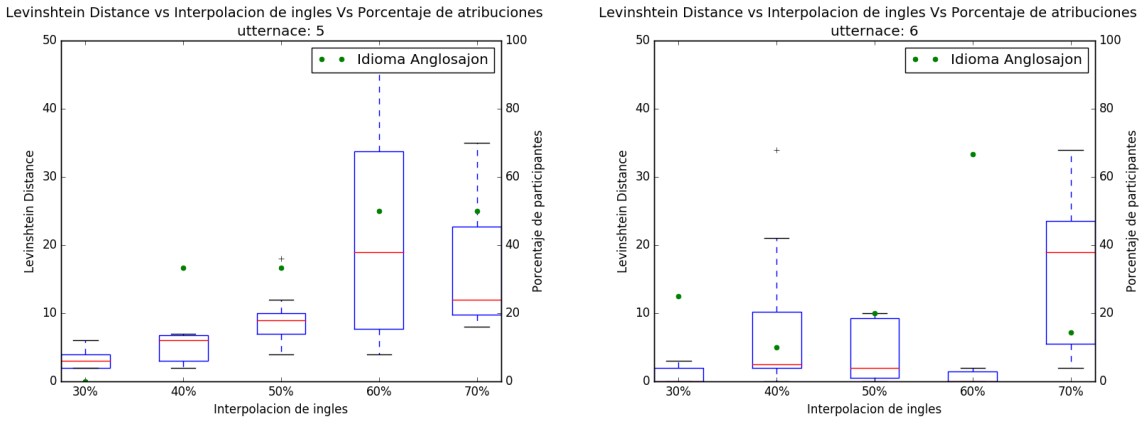


Fig. 3.19: Oración 5 y 6

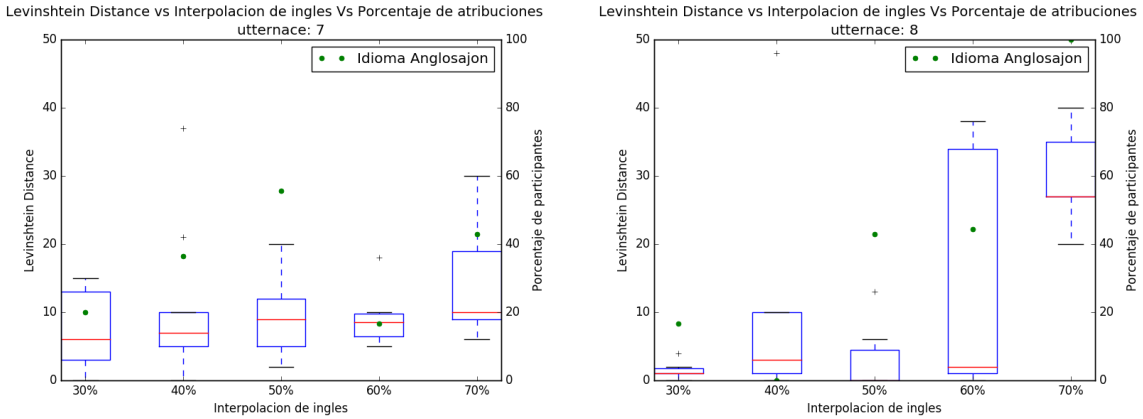


Fig. 3.20: Oración 7 y 8

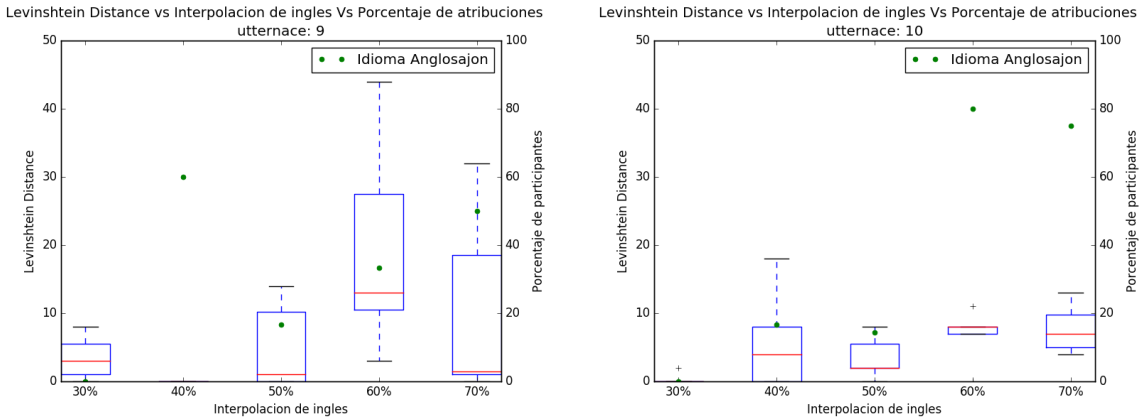


Fig. 3.21: Oración 9 y 10

4. CONCLUCIONES Y 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 problemáticos como el caso de la /r vibrante 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

5.2. Mapeo Fonemas del Ingles-Castellano

5.3. Parametros utilizados para el entrenamiento

Bibliografía

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

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

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

[4] 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

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

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

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

[8] 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

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