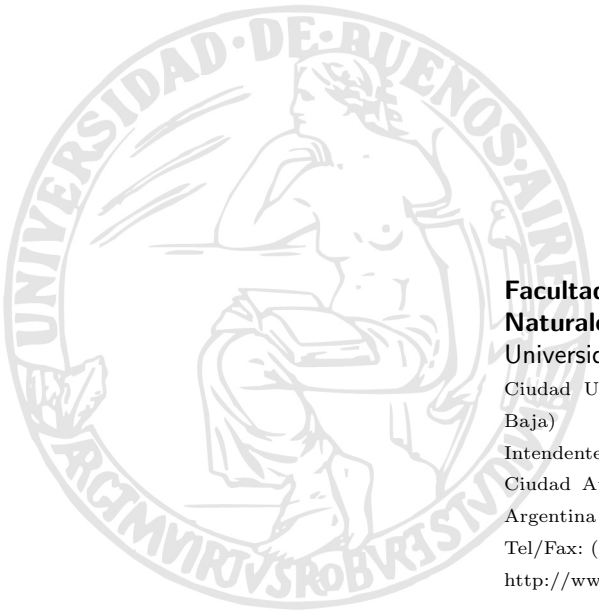
DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

On TTS with non native prosody: a systematic aproach

17 de enero de 2018

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



Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

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

http://www.exactas.uba.ar

1. Introducción

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

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

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

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

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

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

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

2. Metodología

En esta sección presentaremos la metodología utilizada para la generación de HMMs, la interpolación entre los mismos y otras técnicas utilizadas.

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

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

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

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

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

2.1. Preparación De los datos

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

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

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

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

En primera instancia, realizamos varias pruebas de concepto utlizando HTS y este corpus. Para ello fue neceasrio construir los utternaces del mismo. Estos consisten basicamente en una transcripción fonetica de los audios dividida en segmentos temporales y datos contextual tales como la cantidad de silabas en la palabra siendo transcripta, fonemas que preceden y proceden al actual, etc. Estos utternaces serán utilizados en el entrenamiento con HTS para modelar cada uno de los fonemas con una mezcla de variables aleatorias gaussianas.

Para obtener los utternaces se plantearon varias estrategias posibles. La primera a intentar, dada su facilidad, fue utilizando alineamiento automatico utilizando EHMM alignment [4]. Para esto utilizamos Festival, Festvox y las transcripciones grafemicas del corpus. Los resultados preliminares fueron bastante adversos, los audios generados resultaban poco inteligibles notandose claros defectos acusticos en el fonema /rr que sonaba mas como una /r.

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

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

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

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

En este momento del trabajo obtenemos otro corpus de datos $loc1\_pal$ con 1593 oraciones en castellano rioplatense que se aproximan 2 horas y 26 minutos de habla.

Para este corpus no cotabamos con transcripciones foneticas manuales por lo que nos vimos forzados a utilizar EHMM nuevamente. Aún así, los resultados fueron muy superiores a los conseguidos con secyt-mujer. Los audios sintetizados resultaban inteligibles y con un marcado acento rioplatense. Tras ajustar algunos parametros decidimos proseguir con uno de los modelos generados con este corpus.

La teoría de porque hubo tanta disparidad en la calidad de los resultados es cantidad de audios y horas de habla de cada base de datos. Concideramos que esto juega un papel predominante en la calidad de los TTS generados, aún cuando se utiliza un metodo de etiquetado puramente automatico y propenso a errores en el alineamiento.

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

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

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

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

2.2. Repertorio Fonetico y Mapeo De Fonemas

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

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

Por otra parte, surge aquí un problema: como sintetizar oraciones en castellano utilizando un repertorio fonetico distinto, donde incluso la cantidad de fonemas es diferente. Como solución a esto desarrollamos de manera perceptual

e iterativa, un mapeo del ingles al castellano en el que cubriremos cada fonema del castellano por al menos uno del ingles. El mapeo que concideramos devolvió los mejores resultados puede observarse en el apendice.

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

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

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

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

2.3. Entrenamiento Con HTS

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

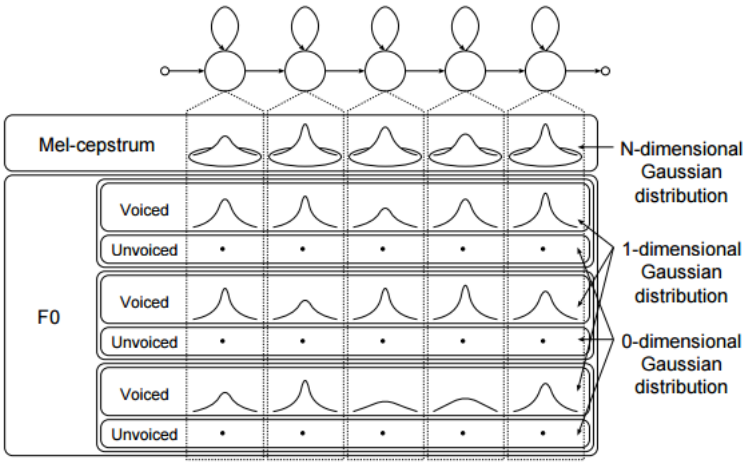


Figure 5.3: structure of HMM.

Por otra parte 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 conseguida en los utternaces. A continuación se presenta una vista esquemática de la estructura de este HMM:

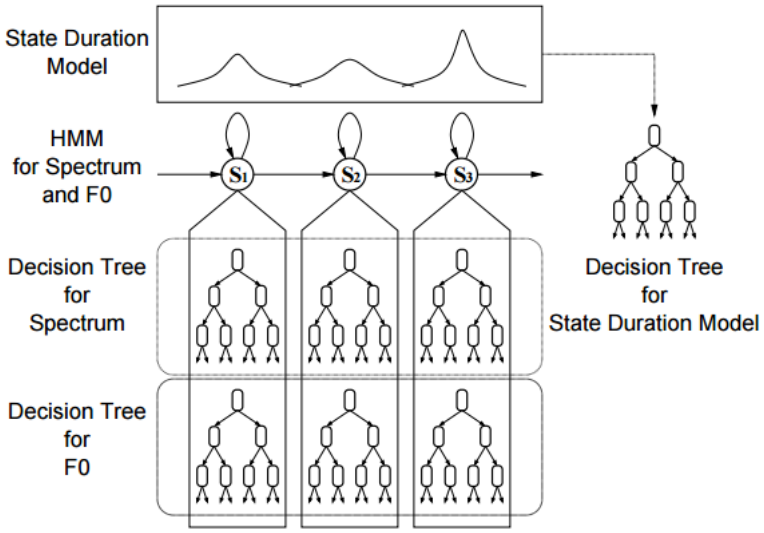


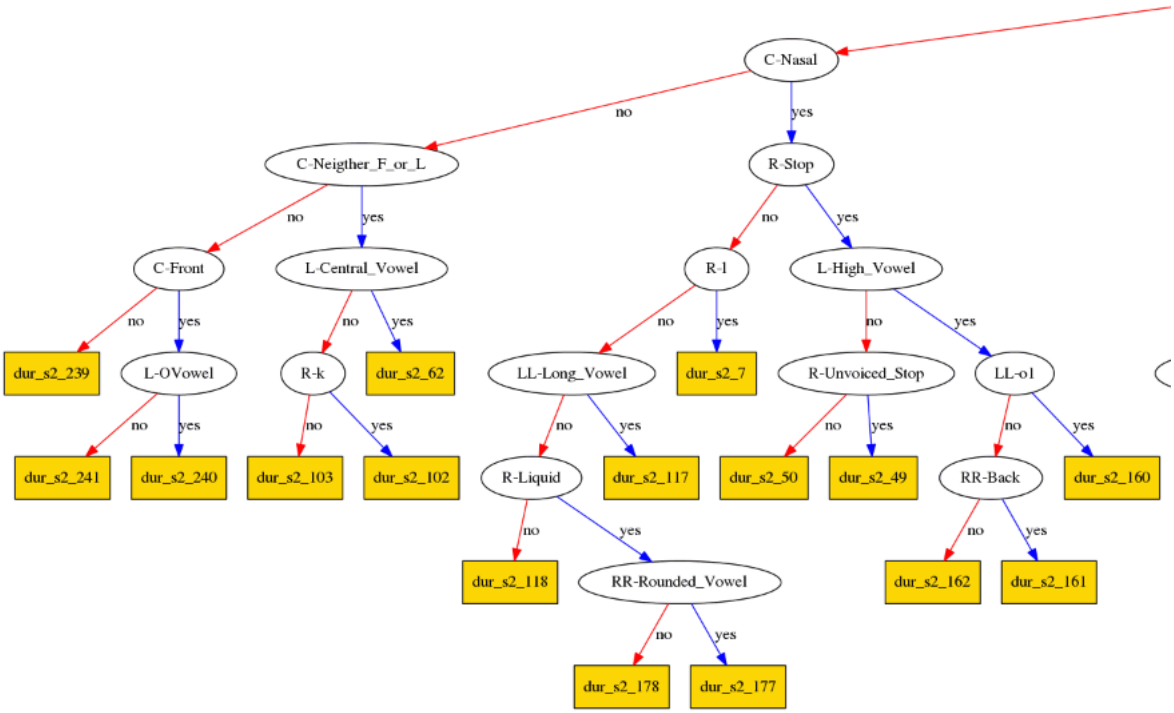
Figure 5.4: Decision trees.

(Aclarar: Imágenes extraídas de la disertación doctoral, Profesor Tadashi Kitamura)

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

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

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



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

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

2.4. Síntesis utilizando hts_engine

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

Comenzamos realizando pequeñas pruebas internas para probar la efectividad del metodo y concluimos que eran satisfactorias, era posible generar oraciones donde la carga fonetica era distintivamente estadounidense (detalles como la /r mas suavizada, o las vocales mas abiertas)

En preparación para la experimentación posterior, surge la pregunta de si es posible utulizar tecnicas de speaker adaptation para...

2.5. Experimentación

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

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

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

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

La misma se llevó a cavo desde el 18 de octubre de 2017 hasta el primero de diciembre.

2.5.1. Interface

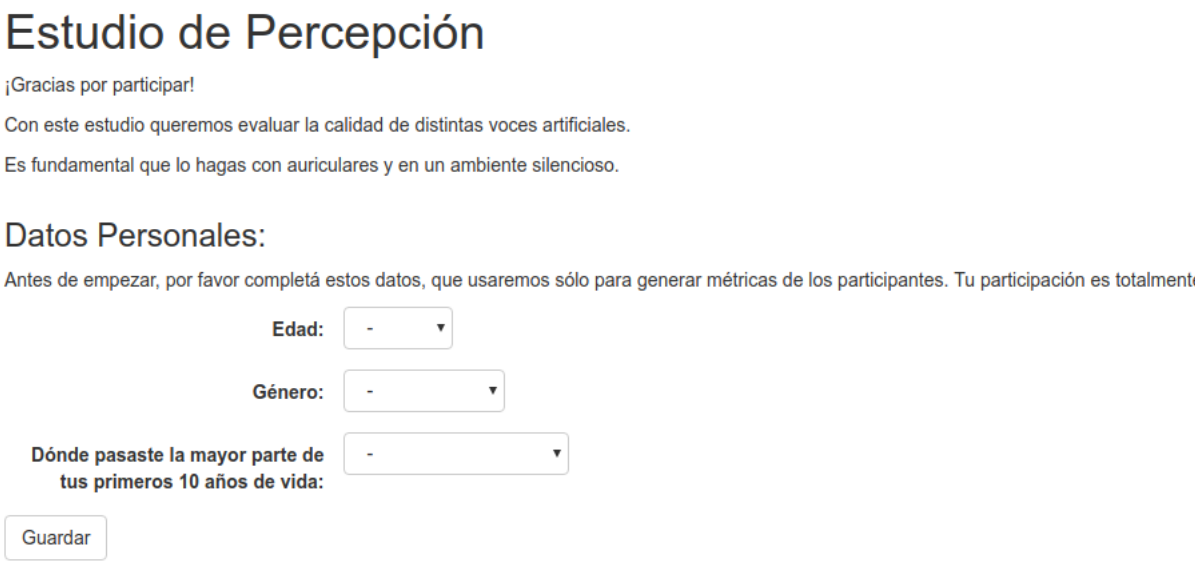
A continuación se presentará la interface que utilizamos para realizar la encuesta junto con las deciciones de diseño que fueron tomadas a lo largo de la misma.

Todos los participantes al entrar en la pagina donde se hosteaba la encuesta se encontraron con lo siguiente:

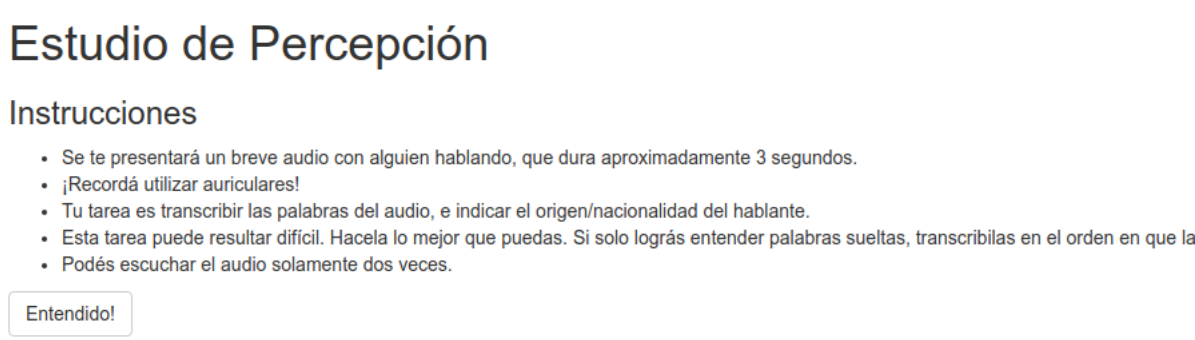
Con el objetivo de no influir el las respuestas de los participantes participantes, procuramos no darles a los participantes en ningun momento información especifica de que buscabamos con este estudio.

A cada participante se le pidió como requerimiento que realizara la encuesta con auriculares y en un lugar silencioso.

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



Una vez que completaban esta información, y presionavan el boton de guardar, se les presentaba otra vista donde se le brindaban las instucciones necesarias para completar la encuesta:



Una vez que precionan el boton de .entendido!"les es presentado un audio, que pueden escuchar un maximo de 2 veces, una caja de texto libre donde escribir lo que interpretaron del mismo y una caja de texto libre donde escribir la nacionalidad que concideran es el hablante.

Las oraciones presentadas fueron:

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

2.6. Resultados

Se encuestaron 109 participantes, con los que se obtuvieron 352 resultados. A continuación presentaremos los datos demograficos de los participantes:

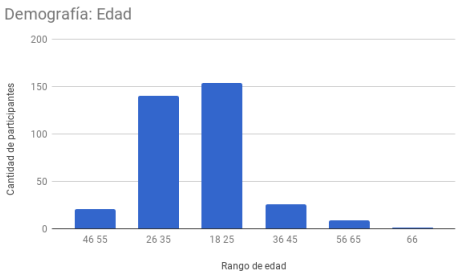


Figura 1: Figure caption

En cuanto al genero de los participantes, 187 respuestas fueron brindadas por participantes del genero femenino mientras que 163 respuestas fueron respondidas por personas del genero masculino, una persona no contesto a esta pregunta.

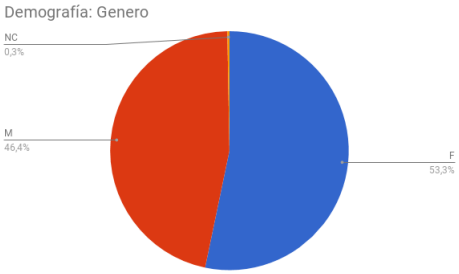


Figura 2: Figure caption

Así mismo, la distribución de el lugar donde los participantes pasaron su infancia se puede ver en este grafico:

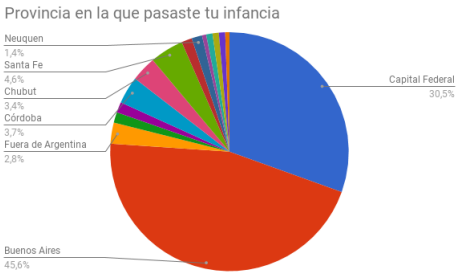


Figura 3: Figure caption

2.6.1. Inteligibilidad

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

Presentamos aquí los resultados obtenidos sin ningun tipo de modificación:

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

Analizando mas detenidamente los datos obtenidos pudimos observar algunas fallas que podrían generar ruido en el analisis, tales como:

- Los participantes escribieron de manera diferente cuando no entendieron un segmento del audio.
  - Muchos de ellos escribieron: "...", ".....º simplemente omitían la palabra.
  - En casos menos comunes: "***", ¿??", "blablabla".

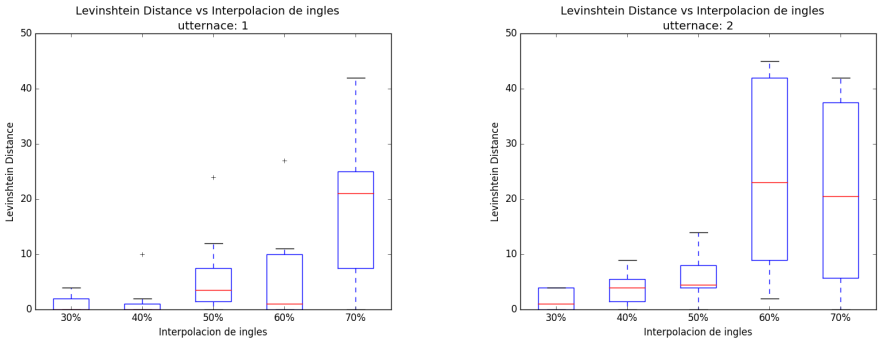


Figura 4: Figure caption

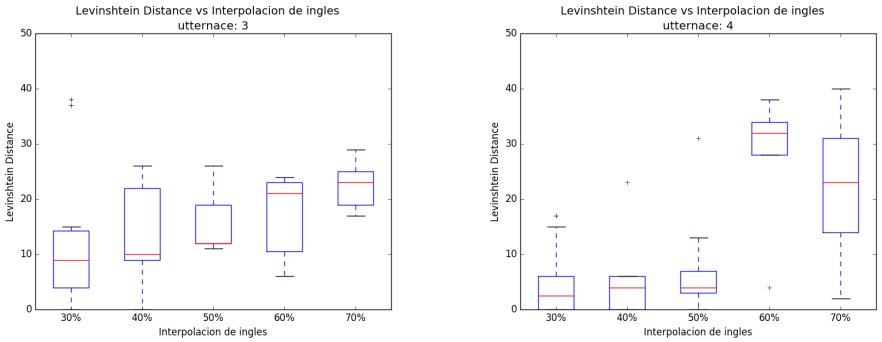


Figura 5: Figure caption

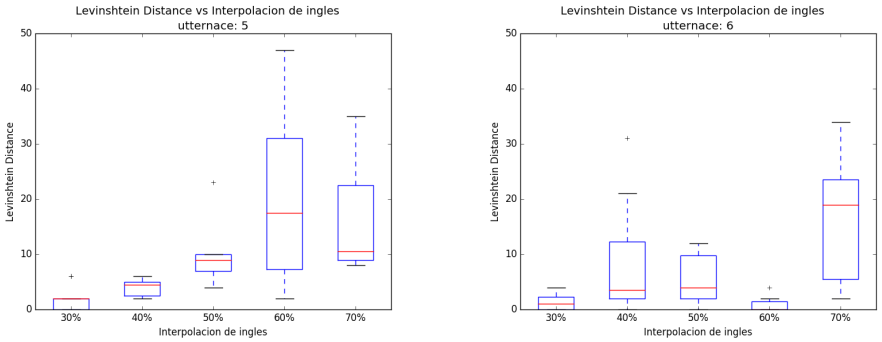


Figura 6: Figure caption

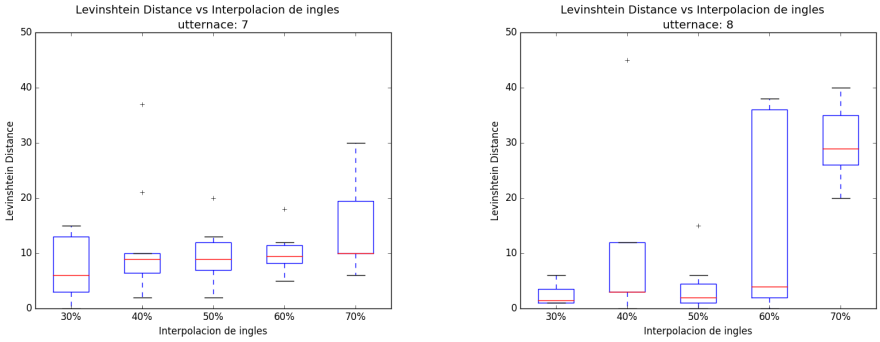


Figura 7: Figure caption

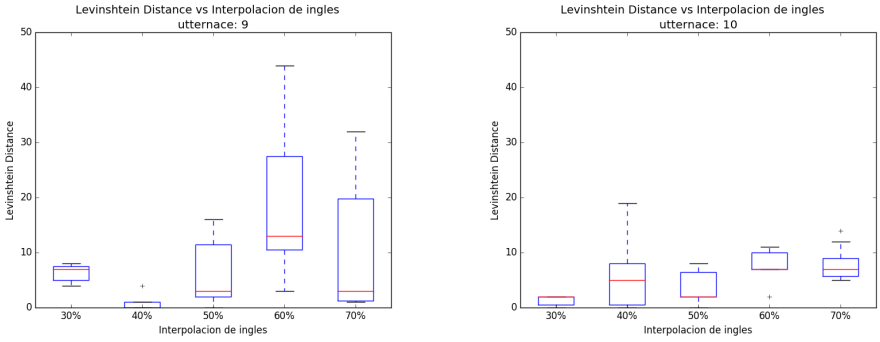


Figura 8: Figure caption

- En casos donde no comprendieron ninguna palabra del audio escribieron cosas como "no entendi nada", "nada", dejaron el campo vacío, etc.
- Utilización de signos de puntuación en las oraciones:
  - Puntos finales para expresar el final de la oración o expresiones como "(?)".
  - En un caso extremo, un participante el participante transcribió tu estrecho posavasos", grito la fechoría", cuando el utternace original solo decía "tu estrecho portavasos gritó la fechoría".
- Omición de acentos en palabras que no resultaban ambiguas.

Para tener datos mas precisos, decidimos realizar una limpieza de los datos donde concideramos que no era disruptiva.
Los cambios fueron:

- corregir "ni"por "ñ.$^{\text{en}}$ la palabra grunion.
- Remoción de todos los signos de puntuación.
- Remoción de expresiones como "blabla.$^{\text{o}}$ cualquier otra que exprese ininteligibilidad de una palabra u oración.
- Correción de acentos en palabras no ambiguas: "botón", "prefirió", recorrió", çhupetín", riñón", "grúñón".
- Las palabras que presentan ambibalencia, como : çoncluyó"no fueron modificadas ya que concluyó/concluyo son validas.

Habiendo realizado estas correcciones, ahora las distancias de Levinshtein se ven así:

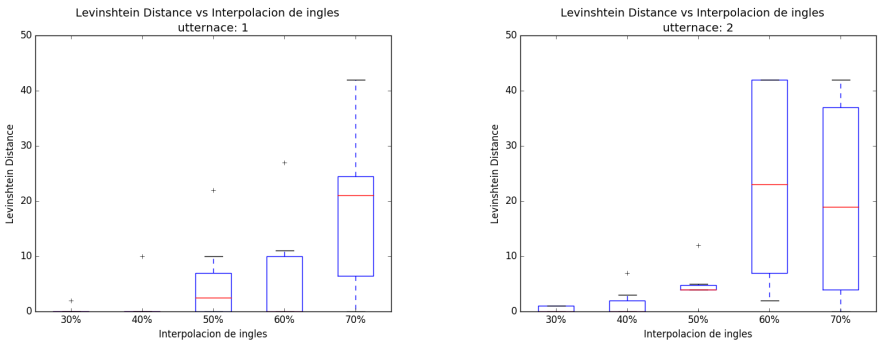


Figura 9: Figure caption

En lo de elegir una nacionalidad: Sinonimos: EEUU-EstadosUnidos, España-España (sur) Cosas no muy especificas: .ªnglo", Ïnglés/Estado Unidense", "Hablante nativo de inglésÇosas raras: "Brasiltiño", ".?", robolandia

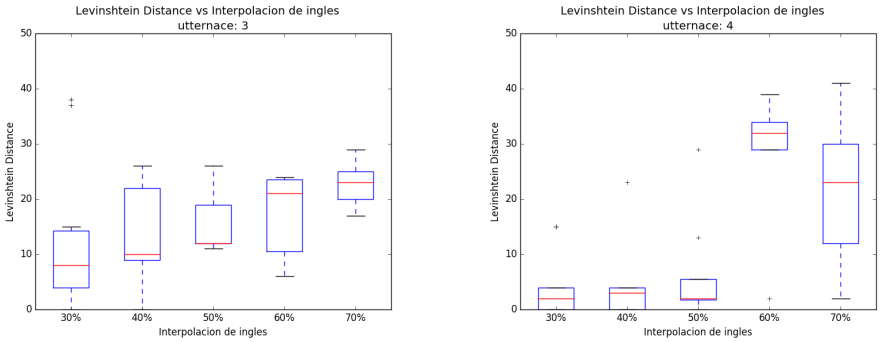


Figura 10: Figure caption

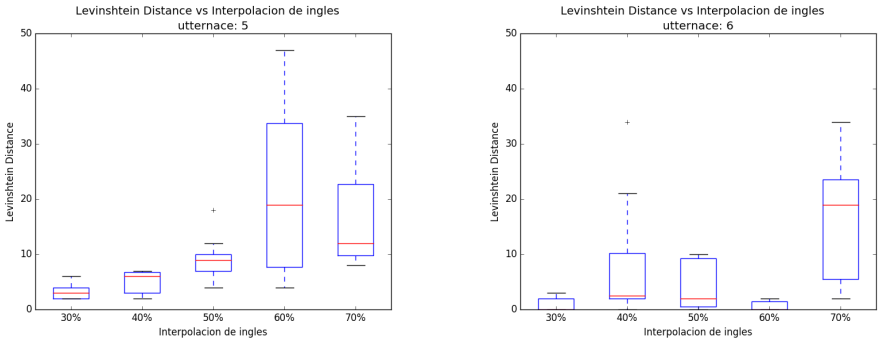


Figura 11: Figure caption

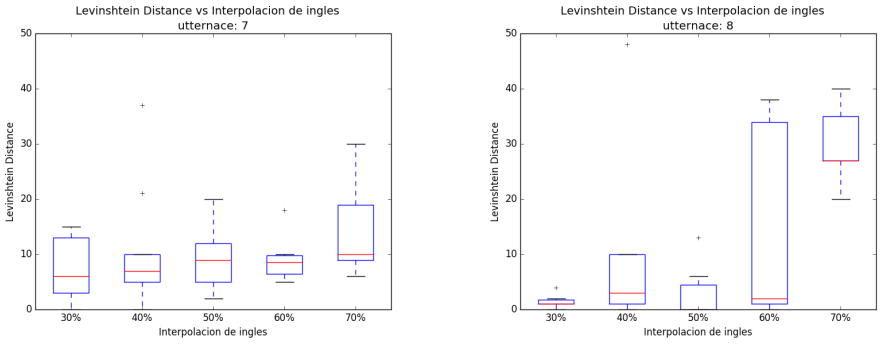


Figura 12: Figure caption

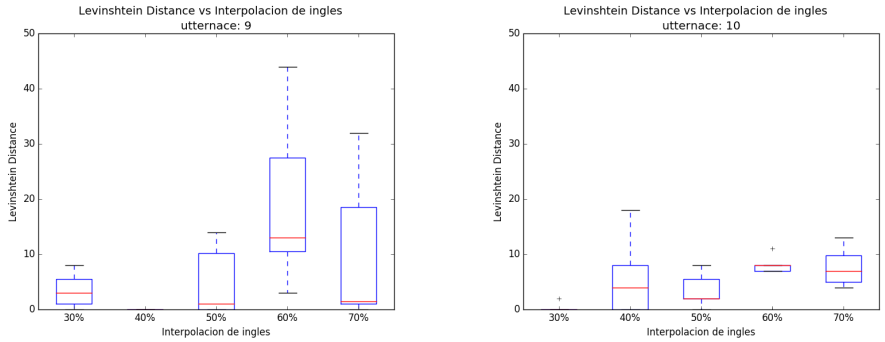


Figura 13: Figure caption

3. Apendice

3.1. Lista de Fonemas

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

3.2. Mapeo Fonemas del Ingles-Castellano

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

3.3. Parametros utilizados para el entrenamiento

4. Referencias

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