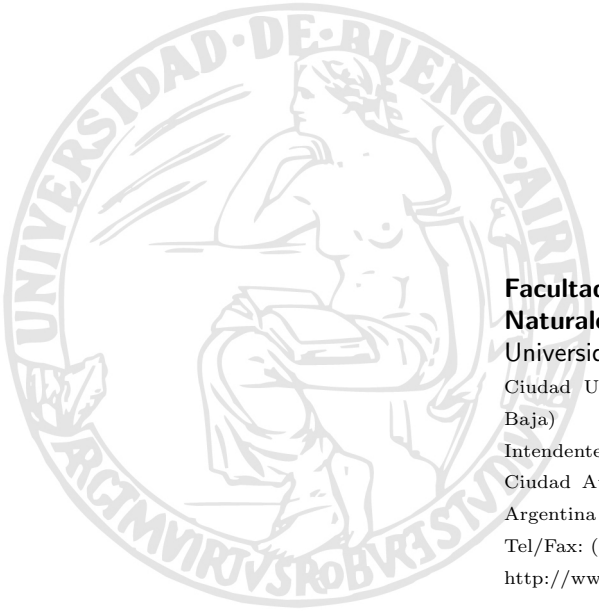
DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

On TTS with non native prosody: a systematic aproach

15 de noviembre de 2017

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.

Para poder realizar el entrenamiento con HTS, fue necesario generar los utternaces respectivos para cada corpus. Los mismos consisten basicamente en una transcripción fonetica de los audios dividida en segmentos temporales y metadata contextual como la cantidad de silabas en la palabra siendo transcripta, fonemas que preceden y proceden al actual, etc. Estos utternaces serán utilizados para modelar cada fonema con una mezcla de variables aleatorias gaussianas.

Dadas la cantidad de horas de audio disponibles, y al costo que conlleva realizar transcripciones manuales de un corpus, tanto para loc1_pal como para CMU-ARCTIC-SLT se decidió utilizar alineamiento forzado automático para obtener los utternaces mencionados. Para esto se utilizaron Festival y Festvox que a partir de los audios y sus transcripciones grafemicas, permite realizar EHMM alignment sobre el corpus de datos. Para secyt-mujer contábamos previamente con las transcripciones fonéticas realizadas a mano, por lo que primero probamos realizar un etiquetado con EHMM alignment, y luego un metodo mixto donde utilizambamos parte de la información del EHMM (principalmente para determinar que fonema correspondía en cada segmento) y las transcripciones a mano para ajustarlos de manera mas presisa.

Comparando resultados, el modelo generado con secyt-mujer y EHMM resulto ser el modelos que peor resultados arrojó, sintetizando oraciones con muchos clicks y voces mas metalicas. El metodo mixto si bien arrojó resultados mejores, todavía presentaba algunos clicks y una voz no del todo natural. Los mejores resultados fueron obtenidos por loc1_pal con EHMM. Nuestra teoría es que esto se debe a la disparidad en la cantidad de audios y horas de habla. 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.

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

2.2. Entrenamiento Con HTS

Tanto para el entrenamiento y síntesis del habla se utilizará HTS.

Comenzaremos dando un resumen del funcionamiento del sistema utilizado:

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

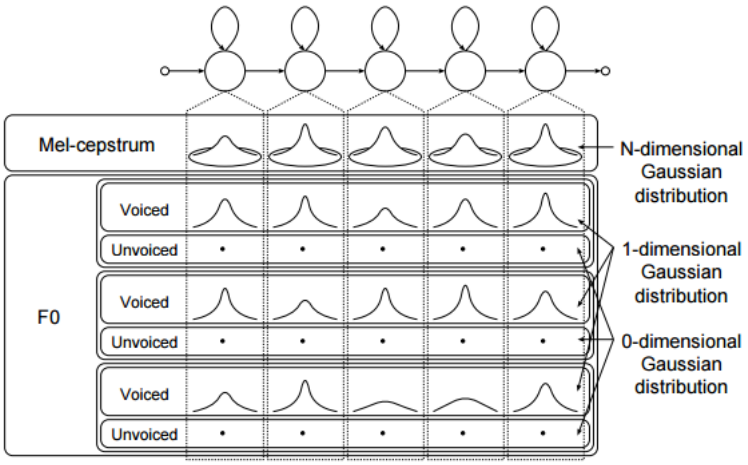


Figure 5.3: structure of HMM.

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 técnicas de aprendizaje automático y arboles de decisión. A continuación se presenta una vista esquemática de la estructura de este HMM:

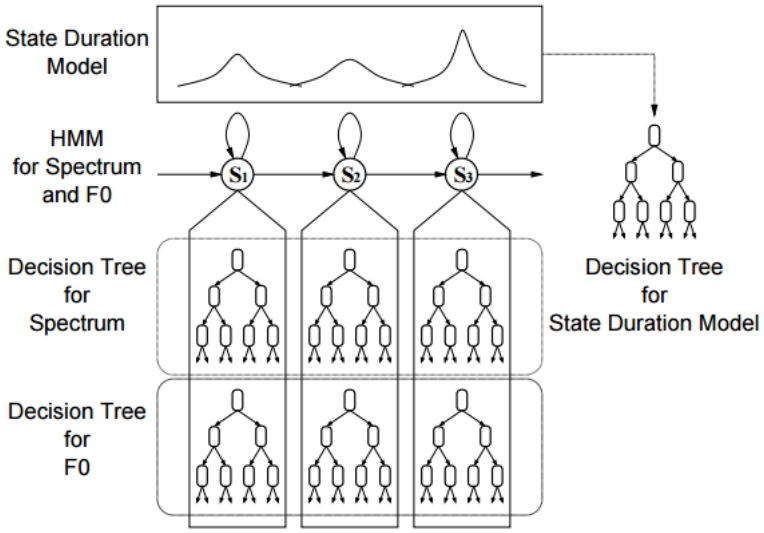


Figure 5.4: Decision trees.

En particular para este trabajo el entrenamiento de todos los modelos se realiza utilizando senones (segmentos de 5 fonemas) para los HMM generados.

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

2.3. 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.

2.4. Síntesis utilizando hts_engine

Para la síntesis se utilizó hts_engine, una herramienta de linea de comandos que no solo permite sintetizar oraciones utilizando los modelos acústicos generados sino además interpolar entre los distintos HMMs disponibles. Utilizaremos esta herramienta para interpolar entre los HMMs de ingles y castellano para lograr nuevos modelos que mezclen los features acústicos con distintos grados de ingles y de castellano.

Un desafío que se presenta para este trabajo es el mapeo de los fonemas del ingles al castellano. Para empezar, la transcripción fonética realizada por festival de las oraciones en ingles puede utilizar 50 símbolos distintos, mientras que la transcripción fonética del castellano utiliza 31. Habiendo además muchos símbolos sin equivalencia. (por ejemplo, con el fonema *rr*).

Para resolver esto desarrollamos una solución adhoc que consistió en desarrollar una función sobreyectiva que permita tener cubiertos los 31fonemas del castellano por alguno del ingles.

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).

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
-	zh

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

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