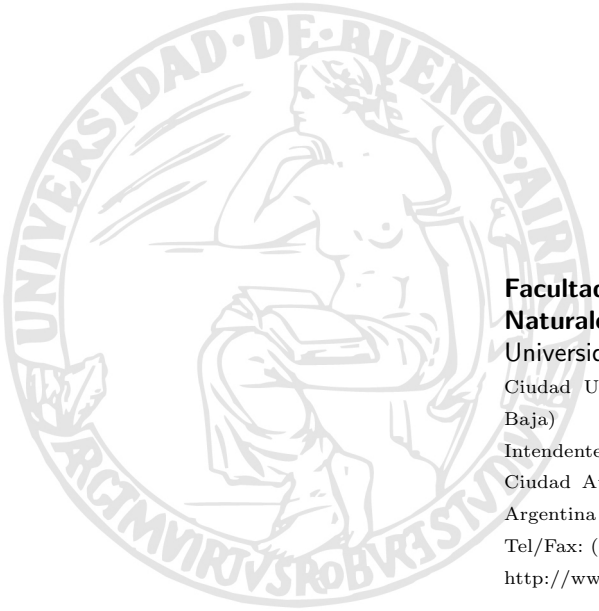
DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

On TTS with non native prosody: a systematic aproach

9 de junio 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. Objetivo

El objetivo de este trabajo se sentrará en explorar diversas tecnicas para sintetizar habla en español con acento extranjero.

Para ello utilizaremos HTS para el entrenamiento y sintesis del habla y Festival y Festvox para realizar el etiquetado automatico de los datos.

Partiendo de tres corpus de datos, dos de ellos en castellano y uno en ingles, generaremos tres HMMs distintos (dos HMMs en castellano y uno en ingles) y utilizando herramientas provistas por HTS interpolaremos entre ellos para obtener distintos grados de acento ingles a la hora de sintetizar.

Dado que el castellano y el ingles no utilizan los mismos simbolos foneticos, una problematica a resolver será mapear fonemas del ingles con su fonema mas similar del cercano.

2. Metodología

2.1. Preparación De los datos

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

- 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.

Dadas la cantidad de horas de audio disponibles, tanto para loc1_pal como para CMU-ARCTIC-SLT decidimos utilizar alineamiento forzado para obtener las transcripciones foneticas necesarias para el entrenamiento. Para esto se utilizó 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 contabamos previamente con las transcripciones foneticas ya realizadas por lo que decidimos utilizar estas.

Por otra parte, festival nos permitirá generar features contextuales sobre cada fonema, como el fonema que lo precede, cantidad de palabras en la oración, si la silaba en la que se encuentra esta acentuada, etc. Mas adelante en este trabajo se explicará de que manera son utilizados estos features.

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

2.2. Entrenamiento Con HTS

Comenzaremos dando un pequeño resumen del funcionamiento del sistema utilizado:

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

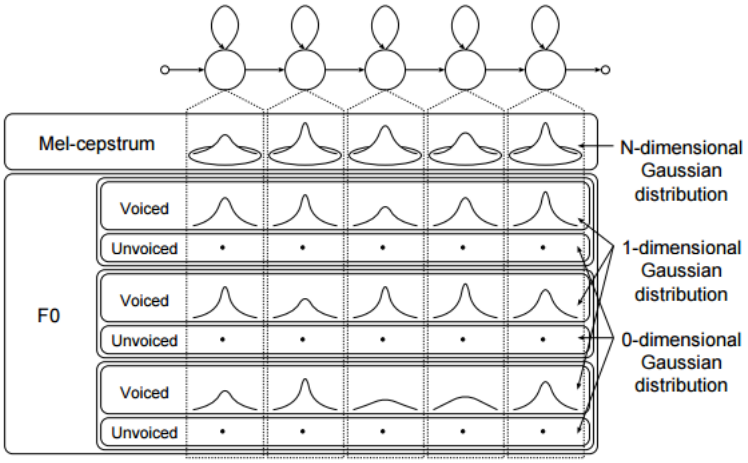


Figure 5.3: structure of HMM.

Por otra parte HTS toma la decición de modelar la información prosodica dentro de este mismo framework. Para esto, las distribuciónes para el espectro, la frecuencia principal y las duraciónes son clusterizadas independientemente utilizando tecnicas de aprendisaje automatico y arboles de decición. A continuación se presenta una vista esquematica de la estructura de este nuevo hmm:

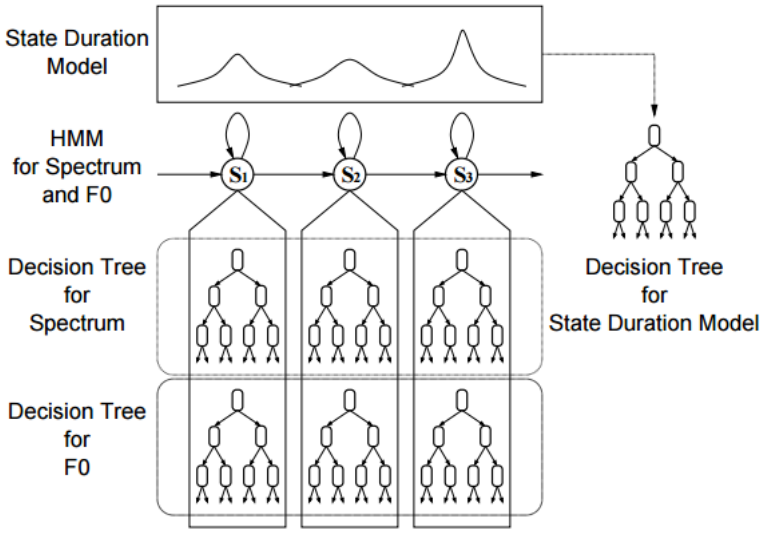


Figure 5.4: Decision trees.

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

2.3. Sintesis utilizando hts_engine

Para la sintesis se utilizó hts_engine, una herramienta de linea de comandos que no solo permite sintetizar oraciones utilizando los modelos acusticos generados sino ademas 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 acusticos 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 transcripcion fonetica realizada por festival de las oraciones en ingles puede utilizar 50 simbolos distintos, mientras que la transcripción fonetica del castellano utiliza 31. Habiendo ademas muchos simbolos sin equivalencia. (por ejemplo, con el fonema /rr).

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