Arturo Romero Blanco

Spanish Emotional Speech Synthesis

School of Electrical Engineering

Project submitted for examination for the degree of
Telecommunication Engineering.
Espoo

Project supervisor:

Prof. Paavo Alku

Project advisor:

D.Sc. (Tech.) Tuomo Raitio

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF
THE FINAL
PROJECT

Author: Arturo Romero Blanco

Title: Spanish Emotional Speech Synthesis

Date: Language: English Number of pages:7+36

Department of Signal Processing and Acoustics

Professorship: Acoustic and audio signal processing Code: S-89

Supervisor: Prof. Paavo Alku

Advisor: D.Sc. (Tech.) Tuomo Raitio

Your abstract in English. Try to keep the abstract short, approximately 100 words should be enough. Abstract explains your research topic, the methods you have used, and the results you obtained.

Keywords: Speech, Emotion,
Synthesis

Preface

This project has been done as part of the Simple4All project in the Deparment of Signal Processing and Acoustics (SPA) at Aalto University School of Electrical Engineering in colaboration with the Deparment of Speech (GTH) of the Technical University of Madrid.
First of all thanks to the GTH group (Juancho, Ruben, Roberto y Jaime) and the SPA group (Paavo, Mikko, Tuomo, Lauri) for the great opportunity and the help. Also thanks to my family for the support and love they have been given to me and to Jose for be patient and stand all my nonsense. And last but not least thanks to Julia for her love, support and for cheer me up in the worst moments.

Otaniemi, 01.04.2014

Arturo Romero Blanco

Contents

List of Figures

List of Tables

Symbols and abbreviations

Symbols

Opetators

Abbreviations

TTS Text to Speech
HMM Hidden Markov Models
HTK
HTS
GlotHMM
MLLRMEAN
CMLLR
MAP
CSMAPLR
SMAPLR
MSD-HMM
HNR
STRAIGHT
MFCCs
PSOLA
GV
ES
MOS

1 Introduction

One of the biggest challenges in speech synthesis is the production of naturally sounding synthetic voices. This means that the resulting voice must be not only of high enough quality but also must be also that it must be able to capture the natural expressiveness imbued in human speech.

Speech synthesis is a field that has been seeing much more use in the last decade with the advent of human-machine interfaces, playing and integral role in them, so applications as telecommunication services, language education, help to people with disabilities, etc can be easily found. As such there have been constant studies on how to improve its quality, naturalness, expressiveness, etc.

Expressive speech synthesis is a sub-field of speech synthesis that has been drawing a lot of attention lately. Assign expressiveness (e.g. emotions [?] or speaking styles) to the synthetic voices will lead to a much more natural voice, increasing the overall satisfaction of the end users of the interface.

Of the two main speech synthesis techniques (unit selection [5] and HMM based) HMM based synthesis has been used in this project due to its parametric nature is much more adaptable. This project is focused in the production of emotional speech synthesis in Spanish language and it is focused in four emotions (anger, happiness, sadness and surprise) plus the neutral voice. This will be done using a text-to-speech (TTS) system, so the input is text with a special format (label) and the system generates the speech waveform. The TTS system is composed by a vocoder (analysis/synthesis tool like STRAIGHT or GlottHMM) and a training module.

This is not the first attempt to do such a thing, it has been tried before and with succeed with the vocoder STRAIGHT (see **??**). So the goal is the use of the vocoder GlottHMM developed in Helsinki, that has been proof to be good in expressive speech recognition [6] and in resynthesis [7] and compare it with STRAIGHT regarding the emotional speech synthesis.

One of the emotional speech synthesis a few years ago was to find a data base with enough data to train a robust model because emotional speech is not easy to find, so it has to be recorded in good conditions and for that money is needed.

The project is organized as follows. The history of the speech synthesis is presented in section **??**, a little bit of emotional theory is explained in section 3, information about the theory used in this project is presented in sections 4 and 7. In section 8 can be found the experiments that have been done and the steps to accomplish them. In section 9 the results of the test performed with the synthesis samples achieved in the experiment can be found and in section 10 the conclusions of this project are exposed.

2 Speech Synthesis History

The earliest successful attempts to produce speech synthesis were made over two hundred years ago [8]. For example, in 1779 by Professor Kratzensteint build some apparatus that represented the human vocal tract to produce five long vowels due to the physiological differences between the vowels. The apparatus were acoustic resonators similar to the human vocal tract and he activated them with reeds like the one used in musical instruments.

The first recorded success in connected speech synthesis was achieved by Wolfgang von Kempelen in 1791 when he completed the construction of his "Acoustic-Mechanical Speech Machine" which was a ingenious pneumatic synthesizer (see figure 1). The machine had a pressure chamber for the lungs, a vibrating reed to act as vocal cords and a leather bag for the vocal tract action. Changing the shape (by hand) of the leather bag different vowel sounds were produced. Constants were simulated by four separate constricted passages that were controlled by the fingers. There were also a couple of hiss whistles to allow the simulation of fricatives and a pair of openings to simulate the nostrils. For plosive sounds a model of a vocal tract that included a hinged tongue and movable lips was employed. To produce a sequence of sounds that seems like speech a lot of practice was needed.

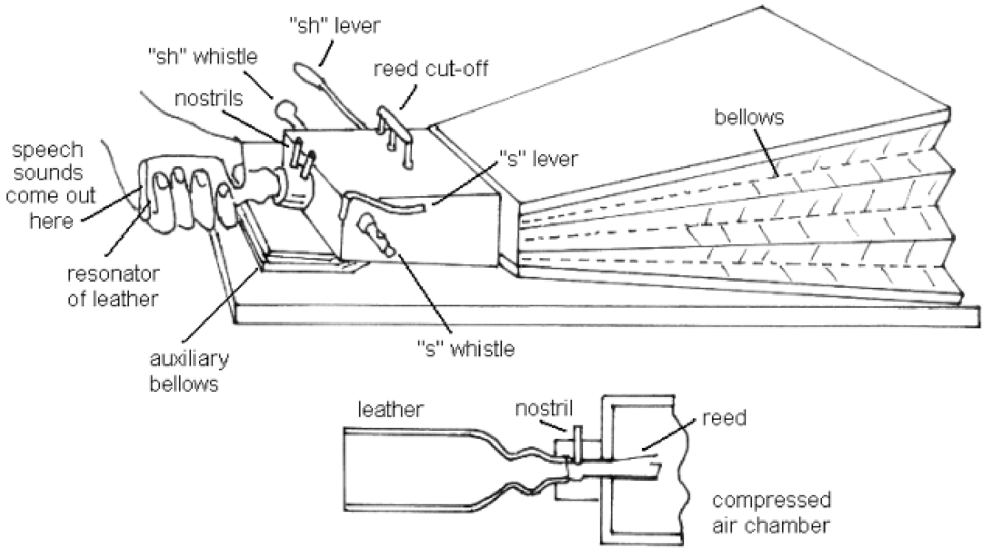

Figure 1: Kempelen Acoustic-Mechanical Speech Machine [1]

The connection between a specific vowel sound and the geometry of the vocal tract was found in 1838 by Willis, who synthesized different vowels with tube resonators and discovered that the quality of the vowel depended only on the length of the tube and not on its diameter. Also in the late 1800's Alexander Graham Bell constructed with his father same kind of speaking machine as the Wheastone's speaking machine that was a reproduction of the Kempelen speaking machine with a few changes.

With the 20th century came the development of electronics and later of electronic resonators. There were a few attempts early in the century to use electronic resonators in such a way that they could produce steady state vowels. An example of this is the electrical synthesis device created by Stewart in 1922. The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowels sounds with two lowest formants, but not any consonants or connected utterances. Obata and Teshima discovered the third formant in vowels. It is considered that the three firs formants are enough for intelligible synthetic speech. It was finally in the late 1930's when the work of Homer Dudley at the Bell Laboratories produced the first electrical connected speech synthesizer.

Dudley developed two devices. One of them, the 'Voder' (figure 2) was basically a parallel array of ten electronic resonators arranged as contiguous band-pass filters spanning the important frequencies of the speech spectrum. It consisted of a wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten band-pass filters whose output gain were controlled via keyboard. A considerable skill was needed to play a sentence on the device and the quality was not good, so in the end it was consider of little practical value, but after the demonstration of the Voder the scientific world became more and more interested in speech synthesis.

The other device Dudley made was called 'channel vocoder'. This channel vocoder and all subsequent vocoders are basically analysis/synthesis devices. They are divided into two halves, an analysis half and a synthesis half. The fist one analyses an incoming speech signal and obtains certain parameters from that natural signal. This parameters are passed as codes to the second half (synthesis) and there they are used to resynthesize a synthetic version of the incoming speech. The channel vocoder is the simplest of the vocoders. It is divided in two branches, one of them determines if the signal is voice or unvoiced and if it voiced it determines the pitch. This information is used to produce a synthetic source. The other branch is a bank of electronic resonators acting like band-pass filters which measure the level of the signal in each frequency band at each point in time. With this information the synthetic source is produced (in the synthesis half of the vocoder) and is mixed with a spectral envelope reconstituted from the filter level values to produce synthetic version of the original signal.

The vocoders were originally developed at the Bell Telephone Labs as devices which allowed a signal to be coded more efficiently and thus allowed more conversations at the same time in the telephone network. More other vocoder configurations have been developed with simply filter banks and rely on complex mathematical transforms of the data (e.g Linear Prediction Coefficient (LPC) vocoders) or on the detection of the formants in the speech signal.

In 1951 the pattern play-back machine (figure 3) was developed by Cooper, Liberman and Borst. It reconverted recorded spectrogram patters into sounds, either original or modified form.

In 1953 Walter Lawrence introduced the first formant synthesizer, PAT, which looked similar to the pattern playback. It consisted of three electronic formant res-

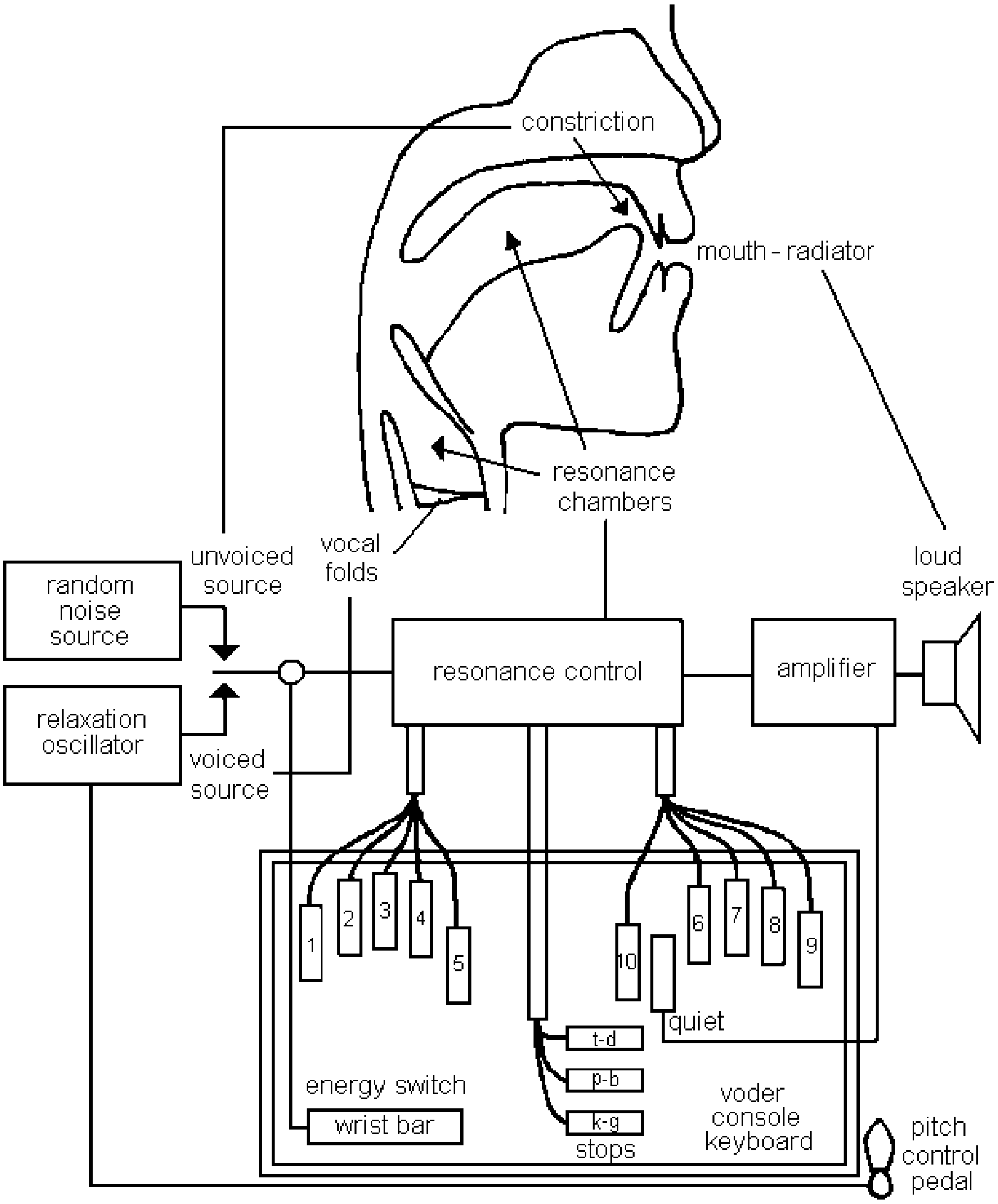

Figure 2: Dudley's Voder speech synthesizer [1]

onators connected in parallel and the input signal was either a buzz or noise. A moving glass slide was used to convert painted patterns into six time functions to control the three formant frequencies, voicing amplitude, fundamental frequency and noise amplitude. At that time Gunnar Fant introduced the first cascade formant synthesizer OVE I which consisted of formant resonators connected in cascade. Ten years later he introduced an improve, OVE II, with Martony which consisted on separated parts to model the transfer function of the vocal tract for vowels, nasal,

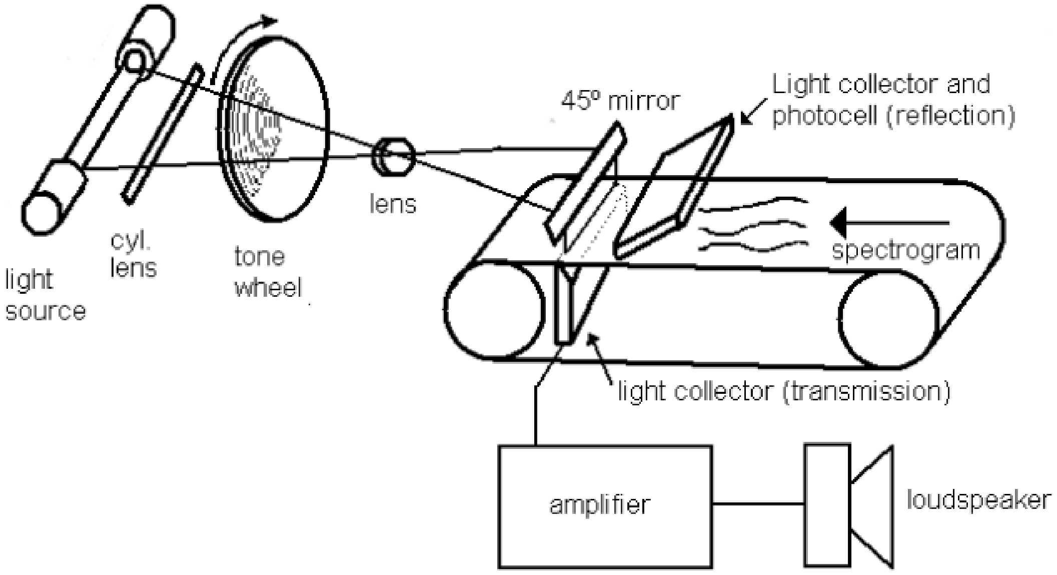

Figure 3: Pattern play-back machine [1]

and obstruent consonants.

In 1958 the first articulatory synthesizer (The DAVO) was introduced at the Massachusetts Institute of Technology by George Rosen. In mid 1960's the first experiments with Linear Predictive Coding (LPC) were made, but it was first used in low-cost systems and its quality was poor. With some modifications this method has been found very useful.

In 1979 Allen, Hunnicutt and Klatt demonstrated the MITalk laboratory text to speech system. Two years later Klatt introduced his Klattalk system, which used a new sophisticated voicing source.

The first reading aid for blind people with an optical scanner was introduced in 1976 by Kurzweil. This system was capable to read quite well multiform written text.

In the late 1970's a lot of commercial TTS and speech synthesis products were introduced. The first integrated circuit was probably the Votrax chip which consisted of cascade formant synthesizer and simple low-pass smoothing circuits. In 1980 The Linear Prediction Coding (LPC) based Speak-n-Spell synthesizer based on low cost linear prediction synthesis chip was introduced by Texas Instruments and it was used for an electronic reading aid for children.

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms. One of the methods applied "recently" in speech synthesis is Hidden Markov Models (HMM, section 4).

3 Emotion Analysis

One of the biggest problems found in research about speech is its variability. The intelligibility of the speech synthesizers is similar to the human one, but they do not have the variability of human speech which makes synthetic voice sound no natural. The emotion is not a simple phenomenon, a lot of factors contribute to this.
Emotions are experienced when something unexpected happens and the emotional effects start to have control in those moments. So emotion can be also described as the interface of the organism with the outside world, pointing three main emotion functions:

- Reflect the evaluation of the importance of a particular excitation in terms of the organism necessities, preferences, etc.
- Prepare physiologic and physically the organism for the appropriate action.
- Notify the state of the organism and its intentions to other organisms that surround it.

Emotion and mood are two different concepts, while emotions happen suddenly in response of a determined excitation and last seconds or minutes, the mood is more ambiguous in its nature and can last hours or days.
A lot of the words used to define emotions and its effects are necessary diffuse and are not clearly defined. This can be explained due to the difficulty for expressing with words abstract concepts that can not be quantified. For that reason, to describe the characteristic of the emotions a group of emotive words are used, but most of them are selected for personal choice.
The first researches about how the emotions affect to the behavior and the language of the animals were briefly described by Darwin in his book *The Expression of Emotion in Man and Animals* [9]. Lately, the effects of the emotions in speech have been studied by acoustic researchers that have analyzed the speech signal, by linguist, that have studied the lexical and prosody effects, and by psychologist . Thanks to them a lot of components present in emotions have been identified. The more important are: pitch, duration and voice quality.
The pitch (F0) is the fundamental frequency at which the vocal cords vibrates. The characteristic of the pitch are some of the main source of information about emotions. For example:

- The average value of F0 express the level of excitation of the speaker, so a high average of F0 means a higher level of excitement.
- The range of F0 is the distance between the maximum and minimum value of the F0. It also reflects the level of excitation of the speaker.
- Fluctuations in F0, defined as the speed of the fluctuation between high and low values and if they are blunt or soft.

The duration is the component of prosody described by the speed of the speech and the situation of the accents, and which effects are the rhythm and the speed. Emotions can be distinguish for some features as:

- Speech speed: usually an excited speaker will reduce the duration of syllables.
- Number of pauses and its duration: an excited speaker will tend to speak faster, with less and shorter pauses, while a depressed speaker will speak slower and with bigger pauses.
- Quotient between speak and pauses time.

The quality of the speech can be distinguish by:

- Intensity: is related with the perception of the volume.
- Voice irregularities: the speech jitter reflects the fluctuations of F0 of a glottal pulse to the other (like in angry emotion) or the disappearance of speech in some emotions (like sadness).
- The quotient between high and low frequencies: a big amount of energy in high frequencies is associated with the angry emotion, while low amount of energy is related with sadness.
- Breathiness and larynx effects reflects the characteristics of the vocal tract that are related with the customization of each voice.

Different classifications have been given to the emotion: The emotions can be divided into primary and secondary emotions [10].

- Primary emotions are those that are considered no acquired through experience but through evolutionary processes. In this group the happiness, sadness, fear, anger, surprise and disgust are found.
- Secondary emotions, are those that derive from previous ones through experience and cognitive modulation.

Joel Davitz and klaus Scherer classified the emotions and its effects using three edges of the semantic field:

- Power or Strength: corresponds to the attention or rejection, differentiating between emotions started by a subject to the ones that appear of the environment.
- Pleasure or evaluation: according to the pleasant or unpleasant of the emotion.
- Activity: presence or absence of energy or tension

Thank to some research it has been discovered that emotions with a same lever of activity are easier to confuse that the ones that have a similar level of strength or pleasure. So the activity is more related with simple hearing variables as tone or intensity.
Some researchers have divided the emotions into two groups, so an emotion can be:

- Active: which qualities are a low speech speed, low volume, low tone and a more resonant timbre.
- Pasive: which qualities are a high speech speed, high volume, high tone and a "turned on" timbre.

Another classification more simple and natural is divide the emotion in positive or negative. Different levels inside this classification can be found.
More information about emotions like biological reasons can be found in [11].

4 HMM

The Hidden Markov Model (HMM) is one the statistical time series model most used in different fields. It has been used in speech recognition for years with great success and also TTS systems has made substantial progress in the last years using HMM.

A HMM is a finite state machine which generates a sequence of discrete time observations. At each time unit, the HMM change the state at Markov process with a state transition probability and the generates observational data in accordance with an output probability distribution of the current state.

A N-state HMM machine is defined by the state transition probability (A), the output probability distribution (B) and initial state probability (Π). Typical HMM structures can be seen in figure 4.

The structure on the left of figure 4 is a 3-state ergodic model, in which all states can be reached by the others in a single transition. The structure on the right is a 3-state left to right model, in which the state index simply increases or stays depending on the time increment. This last model is often used as speech units to model speech parameter sequences since they can appropriately model signals whose properties successfully change.

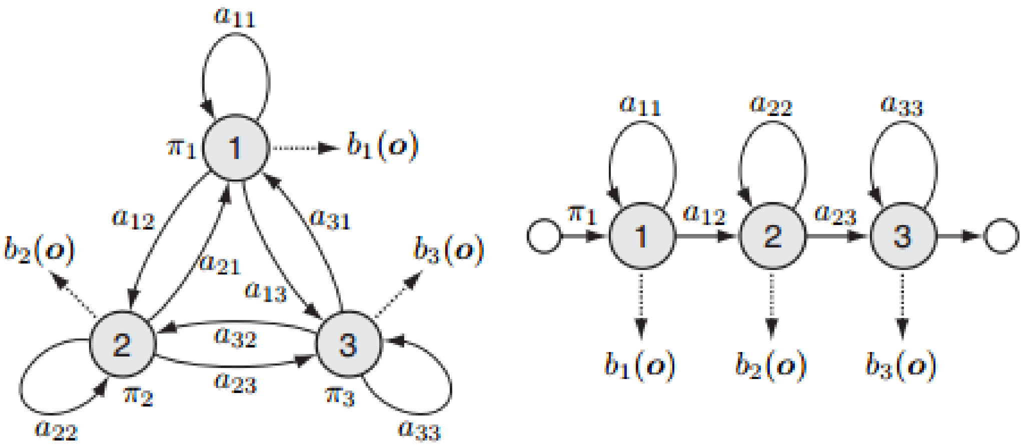

Figure 4: Typical HMM structures [2]

5 HMM-Based Speech Synthesis

Here an HMM-based text-to-speech system is described. In the HMM-based speech synthesis, the speech parameters of a speech unit are statistically modeled and generated using HMMs based on maximum likelihood criterion **??**.

Explicar como funciona un poco ain meternos en el modelo matematico parametros training etc

Luego decir que el HTS puede usar varios vocoder para sintetizar y analizar y hablar ahi de straight (mepcepstral etc) y de glott. A lo mejor buena idea hablar de straight y glott en el analisis para diferenciar un poco training comun y luego volver a hablar de cada uno en synthesis Mejor un poco en general del HMM con training y sinthesis como tiene tuomo y luego explico un poco mas personalizado analsis y synthesis para straight y glott como tiene manu

The main goal of the TTS system is to produce natural synthetic speech sound including different types of speaking and emotions. In order to achieve this the system can be divided into two main parts: training and synthesis, as it is illustrated in figure 5. The analysis is considered as part of the training and is where the features are extracted from the speech database. This features are then modeled by HMM. In the synthesis part, the HMMs are concatenated according to the analyzed input text and speech parameters are generated from the HMM, then the synthesis module transforms them into a speech waveform.

5.1 Training Part

As it has been seen (section 5) this training part is divided into two stages: the parametrization or feature extraction and the HMM training.

In the parametrization stage the input speech signal is compressed into a few parameters which would describe its characteristics as accurately as possible. This stage is done in a different ways depending on the vocoder that is being used and will be explained in section 6 and for more detail see **?? ??**.

In the HMM training stage the features obtained are modeled simultaneously by HMM. First monophone HMM models are trained in a 7-state left-to-right structure with 5 emitting states. All the parameters excluding the F0 are modeled with continuous density HMMs by single Gaussian distributions with diagonal covariance matrix. F0 is modeled with by a multi-space probability distribution (MSD-HMM) [2] due to the conventional continuous or discrete HMMs models can not be applied to F0 pattern modeling because F0 is composed of one-dimension continuous values and a discrete symbol that represents the unvoice. The state duration for each HMM are modeled with multidimensional Gaussian distributions. For GlotHMM (Y NO SE PARA STRAIGHT!!!) each feature is modeled in an individual stream and for the F0 due to the MSD-HMM three streams are used, so the model has eight streams. In order to smooth transitions between states in parameter generation the

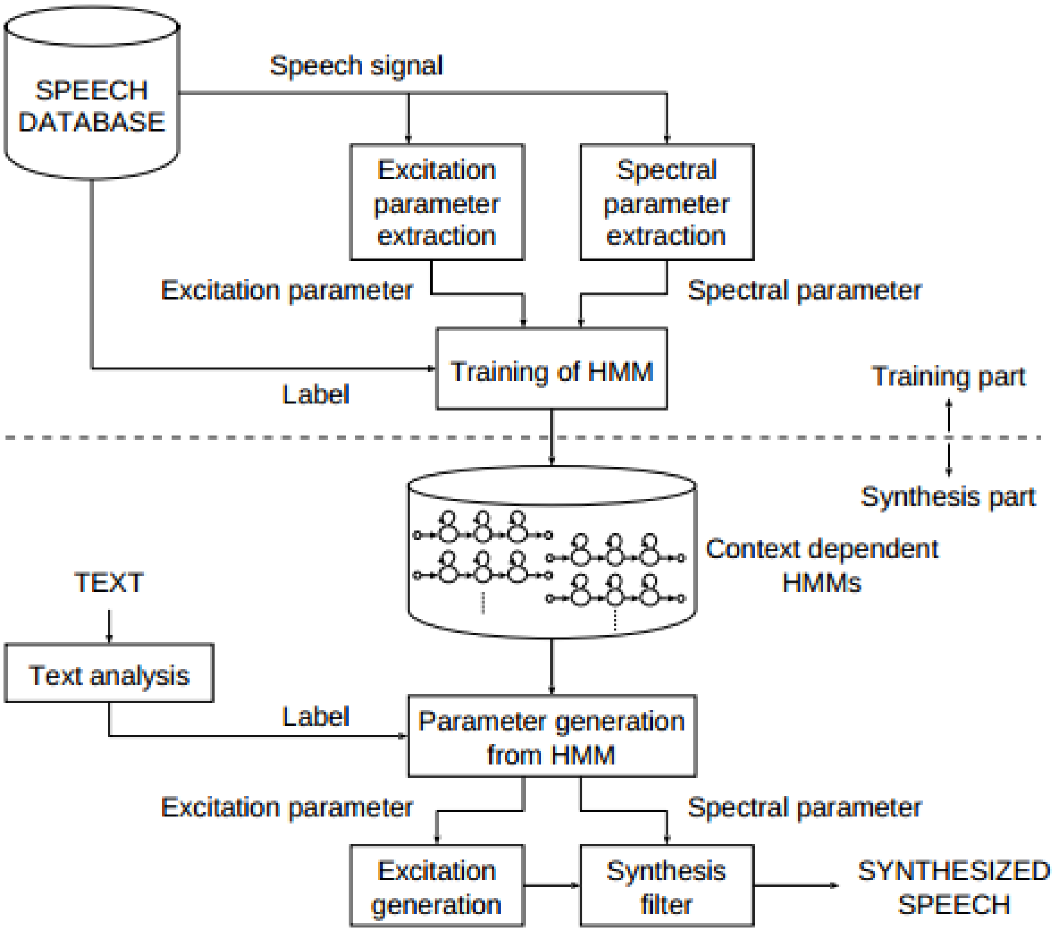

Figure 5: TTS overview [3]

delta and delta-delta coefficients of each feature are calculated, so the total feature order is 171.

After the training of the monophone HMMs the monophone models are converted into context dependent models. As the number of contextual factor increase, their combination increase exponentially. So with limited training data models parameters can not be accurately estimated and it is impossible to prepare a speech database that covers all combinations of contextual factors. To overcome this problem, the models for each feature are clustered independently by using a decision-tree based context clustering (Figure 6). In order to generate synthesis parameters for new observations vectors that are not included in the training data the clustering is also required.

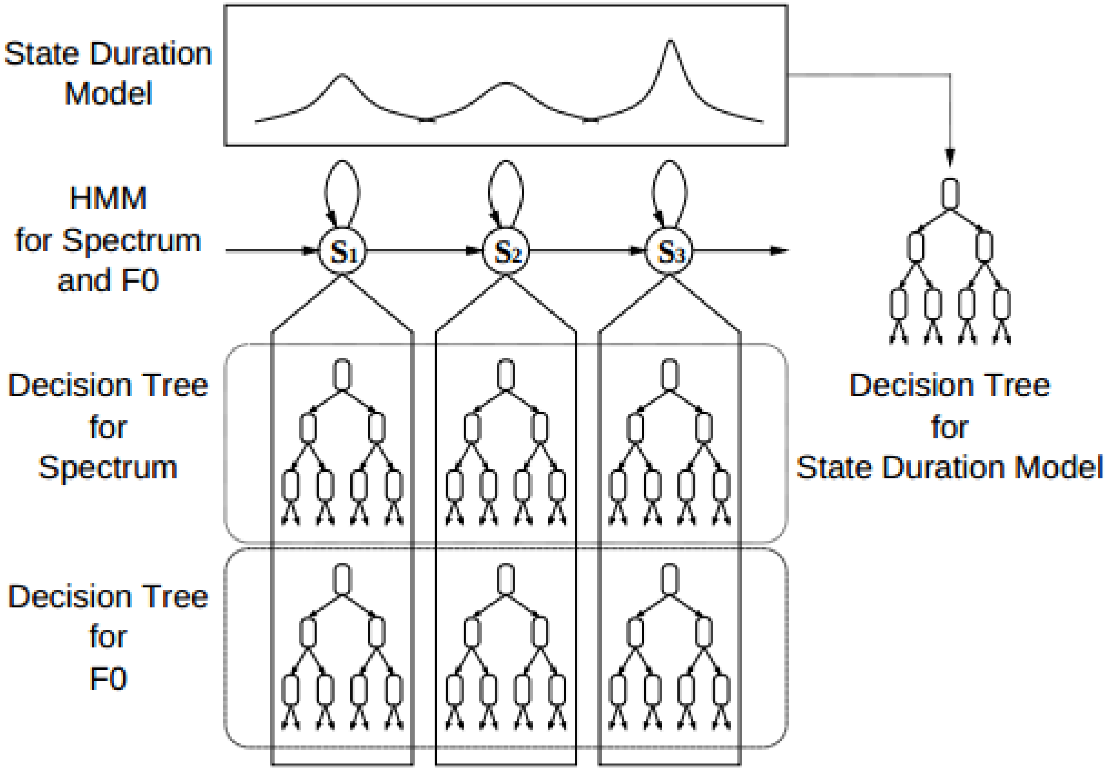

Figure 6: Example of decision-tree based context clustering for some features [3]

5.2 Synthesis Part

In this part the model created in the training part is used to generate speech parameters according to a text input. With this parameters the synthesis module is able to generate a speech waveform. So the synthesis part has two stages: the parameter generation and the synthesis as is illustrated in figure 7.
In the parameter generation stage, the text input is first converted into to a context based label sequence by performing phonological and high level linguistic. According to the decision trees generated in the training stage and the label sequence, a sentence HMM is generated by concatenating the context dependent HMMs. The state durations of the sentence HMM are determined so as to maximize the likelihood of the state duration densities. With the sentence and the state durations, a sequence of speech features are generated and then used by the synthesis module to generate the speech waveform.

In the synthesis stage, as it has already been said, the speech waveform is generated according to the features generated in the first stage of the synthesis part. This synthesis stage differs depending of the vocoder used, so it will be explained in section 6.

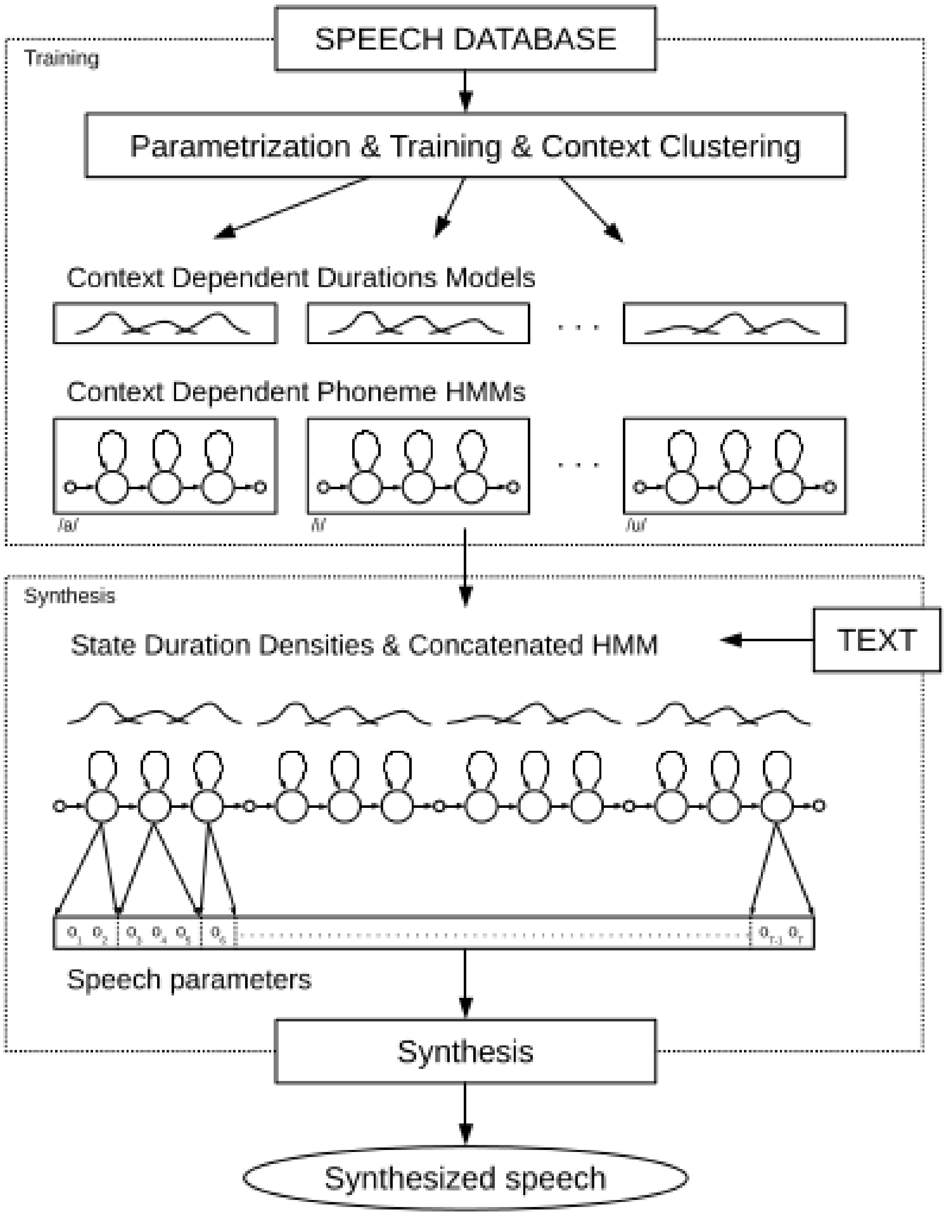
SPEECH DATABASE
Training
Parametrization & Training & Context Clustering
Context Dependent Durations Models
Context Dependent Phoneme HMMs
/a/
/i/
/u/
Synthesis
State Duration Densities & Concatenated HMM
TEXT
Speech parameters
Synthesis
Synthesized speech

Figure 7: HMM-based generation process of speech parameters [4]

6 Vocoders

Many different vocoders has been developed to be applied with HMM-based speech synthesis (see **??**). In this section two of them will be explained: GlotHMM and Straight due to they are the ones that are being compared in this project.

6.1 GlotHMM

The GlotHMM was proposed by Tuomo Raitio [4]. GlotHMM estimates the real glottal pulse signal G(z) an the vocal tract filter V(z) associated with it. So the speech signal can be represented as:

$$S(z) = G(z)V(z)L(z) \tag{1}$$

where L(z) represents the lip radiation. All parts are estimated of real physical properties. For example the glottal pulse signal can be divided into the source part E(z) an the filter containing the spectral envelope of the glottal pulse F_G(z):

$$G(z) = F_G(z)E(z) \tag{2}$$

and so the vocal tract filter can be expressed as:

$$V(z) = \frac{F(z)}{F_G(z)L(z)} \tag{3}$$

To extract the parameters (analysis) of the speech signal GlottHMM follows this steps:

- First, the speech signal is high-pass filtered and windowed into fixed length rectangular frames, from which the signal log energy is calculated as a feature parameter

- Second, the Iterative Adaptive Inverse Filtering (IAIF) algorithm illustrated in figure 8 and explained in **??**, is applied to each frame and results in the LPC representation of the vocal tract spectrum and and the waveform representation of the voice source

- The LPC spectral envelope estimate of the voice source is calculated , and along with the LPC estimate of the vocal tract spectral envelope, is converted into LSF representation

- The glottal flow waveform is used also for the acquisition of the F0 value as well as the Harmonic-to-Noise Ratio (HNR) values for a predetermined amount of sub-bands frequency.

The output of the IAIF algorithm g(n) (estimated glottal flow signal) is used to generate the rest of the analysis parameters. A voicing decision is made based on the amount of zero-crossing and low-band energy. For voiced frames, the F0 value

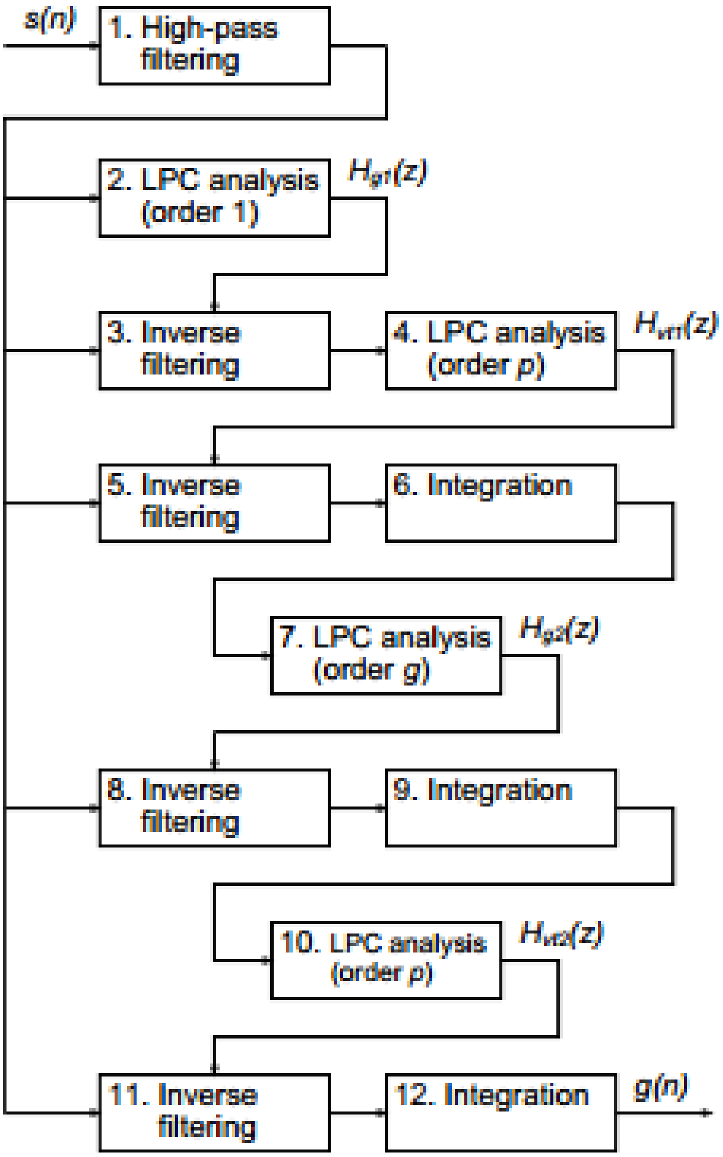

Figure 8: IAIF algorithm block diagram [4]

of the frame is estimated using the autocorrelation method. The HNR is calculated from g(n). For unvoiced frames the HNR and F0 are set to zero. The F0, HNR and the source LSF are used to model the excitation signal that is filtered by the vocal tract filter.

The final analysis vector of GlotHMM consists of single parameters for the F0 and log energy, around 5 parameters for HNR, 10-20 parameters for the glottal source LSF parameters and 20-30 parameters for the vocal tract LSF parameters.

To perform the synthesis GlottHMM uses a method for the excitation generation based on the voice/unvoice decision instead of using a traditional mixed excitation model. The synthesis block diagram is illustrated in figure 9.

For the voiced frames, the heart of the synthesis procedure is a fixed library pulse

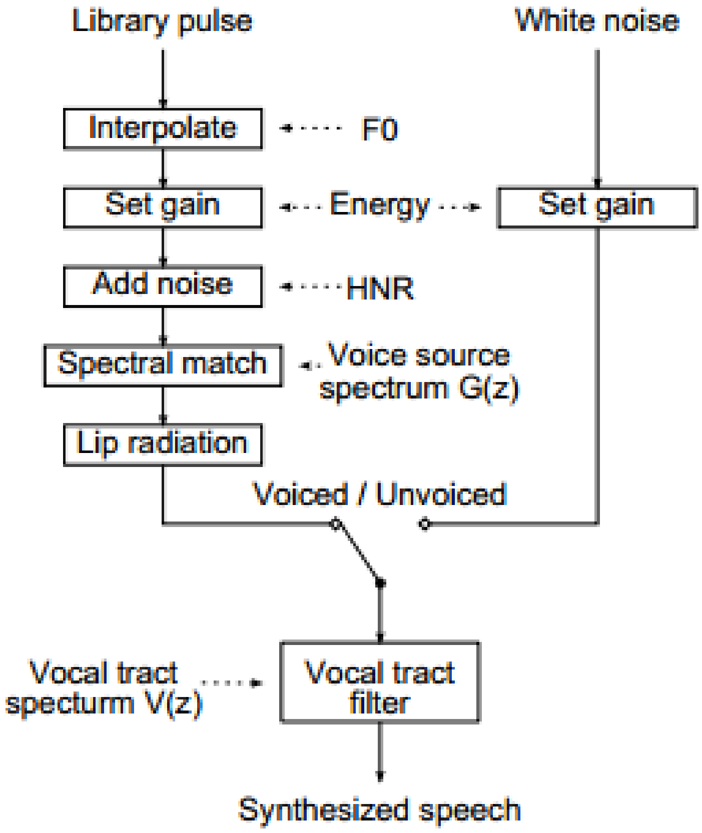

Figure 9: Synthesis block diagram for GlotHMM vocoder [4]

that is obtained by glottal inverse filtering a sustained vowel signal. The library pulse is interpolated to match the target F0 value using cubic spline interpolation, and its energy is set to match the target gain obtained from the analysis vector.

Next, a HNR analysis is done to the library pulse. For each sub-band, noise is added to the real an imaginary parts of the FFT vector according to the differences between the obtained and the target HNR values.

The spectrum of the library pulse is matched to the spectrum of the target glottal pulse obtained from the analysis vector. The spectral matching is done by performing LPC analysis to the library pulse, and then filtering the obtained residual with the target synthesis filter. Finally, the lip radiation effect is added to the excitation by filtering it with a fixed differentiator.

For unvoiced frames, the excitation is generated as white Gaussian noise whose gain is set by the energy parameter of the analysis vector.

The excitation is combined in the time domain by overlap-adding target frames, and the final synthetic signal is generated by filtering the excitation with the vocal tract filter derived from the vocal tract LSFs obtained from the analysis vector.

6.2 STRAIGHT

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) is the more established of the more sophisticated vocoding methods. Proposed by Kawahara in 1977, it has gone through extensive research and development since then. Is often the main reference to which other vocoders in HMM-based synthesis are compared, like in the case of this project.

For HMM synthesis some modifications were made and now the spectral envelope is represented as mel-frequency cepstral coefficients, and the corresponding aperiodicity measurements are averaged over five sub-bands frequency.

In the parameter extraction (analysis) the main idea behind STRAIGHT is the extraction of a smoothed spectral envelope , which minimized the effect of periodicity interference in the analysis frames. This means that the spectral envelope is essentially independent of the speech excitation, which is a great feature with respect to speech transformation.

The extraction of the spectral envelope can be found in **??**.

The spectrum is represented as mel-frequency cepstral representation for the purpose of statistical modeling. The aperiodicity measurements are also transformed into a compressed representation.

The acquired analysis vector for STRAIGHT consists of the F0 value, five aperiodicity coefficients and 20-40 spectral MFC coefficients (MFCCs).

STRAIGHT synthesis is done in frame-by-frame basis by creating a mixed excitation signal of the length of two pulse periods based on the F0 and aperiodicity measurements. The harmonic pulse train is all-pass filtered with a randomized group-delay filter, which reduces the buzziness of the resultant synthesis. The acquired mixed excitation signal is convolved with the minimum phase MLSA filter derived from the frame's spectral MFCCs. Finally, the Pitch-Synchronous Overlap-Add (PSOLA) algorithm is applied to the synthesized frames to get the speech waveform signal. As illustrated in figure 10 the components for the mixed excitation are generated by sub-band filtering the voice (impulse train) and unvoice (white Gaussian noise) parts separately in the frequency domain. The stepwise band-pass filters used are determined by the aperiodicity coefficients so that the resultant sub-bands will have the same average lower-to-upper envelope ratio as the respective aperiodicity coefficient.

After the sub-band weighting, the pulse train component is all-pass filtered to adjust the phase characteristics of the excitation.

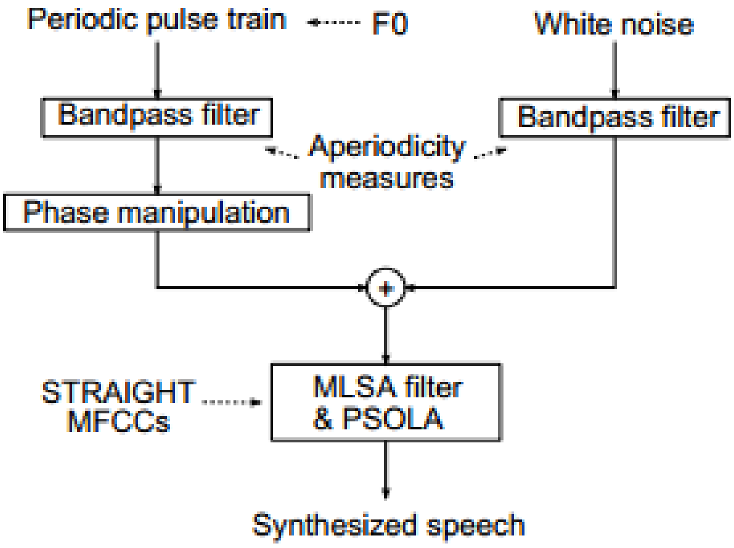

Figure 10: Synthesis block diagram for STRAIGHT vocoder [4]

7 Adaptation

There are several styles of adaptation which affect both the possible application and the method of implementation. Firstly adaptation can be supervised in which case accurate transcriptions are available for all the adaptation data, or unsupervised in which case the required transcriptions must be hypothesis. Secondly, adaptation can be incremental, where adaptation data becomes available in stages or batch-mode, where all of the adaptation data is available from the start.
For cases where the adaptation data is limited, linear transform based schemes are currently the most effective form of adaptation. These approaches use the acoustic model parameters and require a transcription of the adaptation data.

7.1 Maximum Likelihood linear Regression

In maximum likelihood linear regression (MLLR), a set of linear transformations are used to map and existing model set such that the likelihood of the adaptation data is maximized.
There are two main variants of MLLR:

- Unconstrained MLLR: where separate transforms are trained for the means and variances

- Constrained MLLR (CMLLR): where the transform for the mean and the variance is the same

CMLLR is the form of linear transform most often used for adaptive training even with little amount of adaptation data(**??**). For both forms of linear transformation, the matrix transformation may be full, block-diagonal, or diagonal.

7.2 Parameter Estimation

The linear transforms seen in section 7.1 require transcription of the adaptation data in order to estimate the model parameters. For supervised adaptation, the transcription is known and may be directly used without further consideration. When used in unsupervised mode, the transcription must be derived from the recognizer output............YO CREO QUE ESTO NO LO VOY A PONER PORQUE ES MAS PARA RECONOCER Y NO TENGO MUCHA INFO

7.3 Regression Class Trees

A powerful feature of linear transform-based adaptation is that it allows all the acoustic models to be adapted using a variable number of transforms. When the amount of data is limited, a global transform can be shared across all Gaussians in the system, but as the amount of data increases, the HMM state components can be grouped into regression classes with each class having its own transform.

The number of transforms to use for any specific set of adaptation data can be

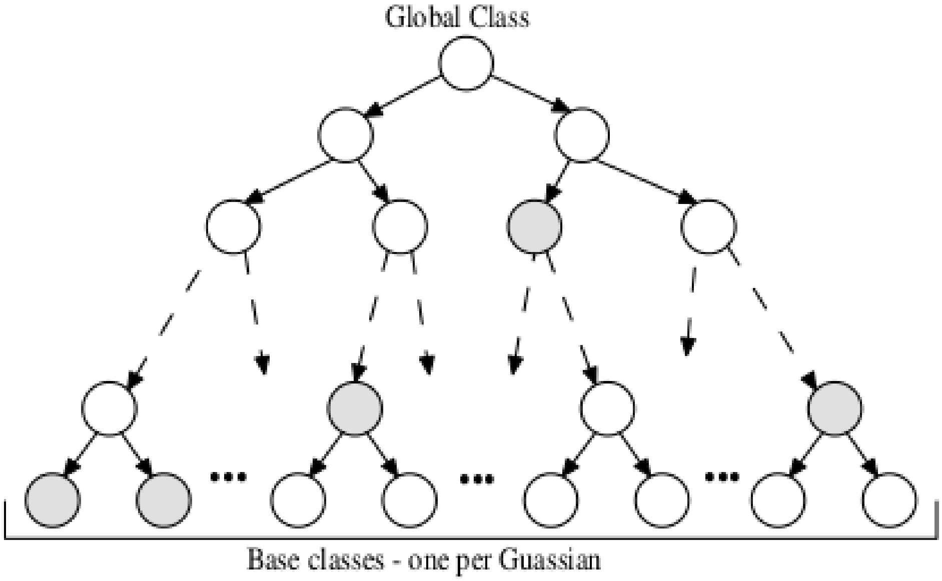

Figure 11: Regression class tree example

determined automatically using regression class trees as illustrated in figure 11. Each node represents a regression class (a set of Gaussian components that will share a single transform). Then, for the given set of adaptation data, the tree is descended and the most specific set of nodes is selected for which for which there is enough data.

7.4 Maximum a Posteriori

Rather than hypothesizing a form of transformation to represent the differences between speakers, it is possible to use standard statistical approaches to obtain robust parameter estimates. One common approach is maximum a posteriori (MAP) adaptation where in addition to the adaptation data, a prior over the model parameters is used to estimate the model parameters.
MAP adaptation effectively interpolates the original prior parameter values with those that would be obtained from the adaptation data alone. As the amount of adaptation data increases, the parameters tend asymptotically to the adaptation domain.

7.5 Adaptive Training

In the case of speaker independent, the training data includes large number of speakers. Hence, acoustic model trained directly on this set "waste" a large number of parameters encoding the variability between speakers rather than the variability between spoken words which is the true aim.One approach to this is to use adaptation transforms during training. This is known as speaker adaptive training (SAT). As

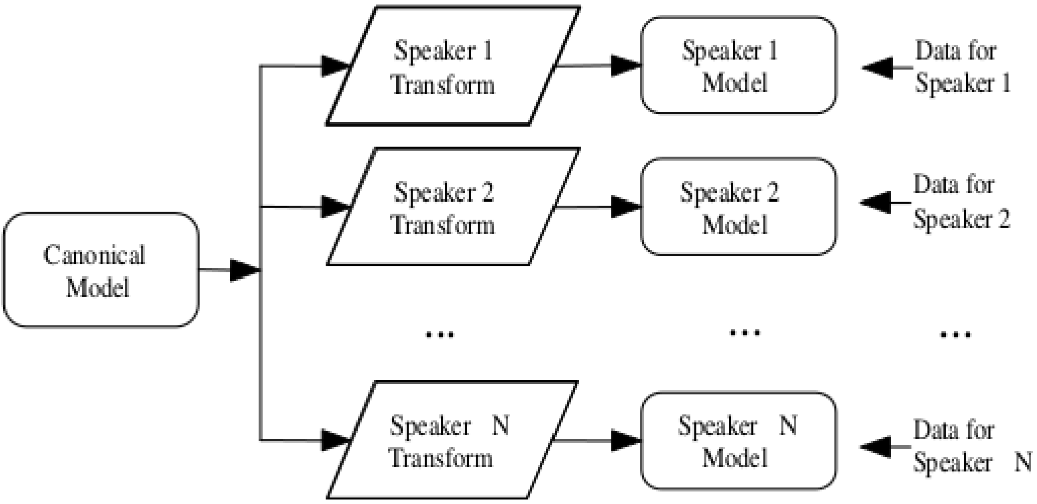

Figure 12: Speaker adaptive training example

example of this is illustrated in figure 12. For each training speaker a transform is estimated and then the canonical model is estimated given all of these speaker transforms. The complexity of this method depend of the nature of the adaptation transform that can be split in three groups:

- Model independent: These schemes do not make explicit use of any model information

- Feature transformation: These transforms also act on the features but are derived, normally using ML estimation, using the current estimate of the model set

- Model transformation: The model parameters, mean and possibly variances, are transformed.

The most common version of adaptive training uses CMLLR, since it is the simplest to implement.

8 Experiments

In this section, the work that has been done will be explained.
The experiments has been carried out with both male and female voices, and two different methods have been applied (Dependent models, section 8.1, Adaptation, section 8.2) in order to get the synthesized voices (section 8.3).
Rergarding the vocoder, GlottHMM has been used and then the results have been compared with STRAIGHT vocoder (see 9).

8.1 Depenent models

The first step in this project was to use dependent models for each emotion, but before some signal processing needed to be done, like sampling the audio files from 44KHz to 16KHz. Once this is done the process for building the dependent models can be started.
In order to get this models the next step were followed:

- Adjust GlotHMM configuration file (8.1.1)
- Adjust HTS configuration file (8.1.2)
- Extract features of the audio files (8.1.3)
- Train the voice (8.1.4)

8.1.1 Configuration File

GlotHMM use a configuration file to extract the features of an audio file (see 6.1), and it is very important have a good configuration to obtain good results before the training.
To try the configuration file, what it is done is to extract the features of a file and then synthesize it without any training done, this is usually called resynthesis. The result of the synthesized file must be very similar to the original one.
A configuration file has been created for each emotion and for some of them the result was better than with others, that is the reason why not all the emotions have the same quality. For example, with the anger emotions the synthesis file is not as similar to the original as the sad one, and this is reflected in the final quality if the voice. This is illustrated in figures 13, 14
Looking into the different configuration files for the different emotions, some little changes can be seen between them. This changes are in the f0 estimation of the analysis, were part the emotion is located (see 3). The rest of the configuration file is the same for all the emotions and it can be also find the parameters that can be extracted in the analysis (can be true or false for all of them) or the ones that will be used in the synthesis. An example of a configuration file can be found in section ??.
In this f0 estimation some values can be tuned as:

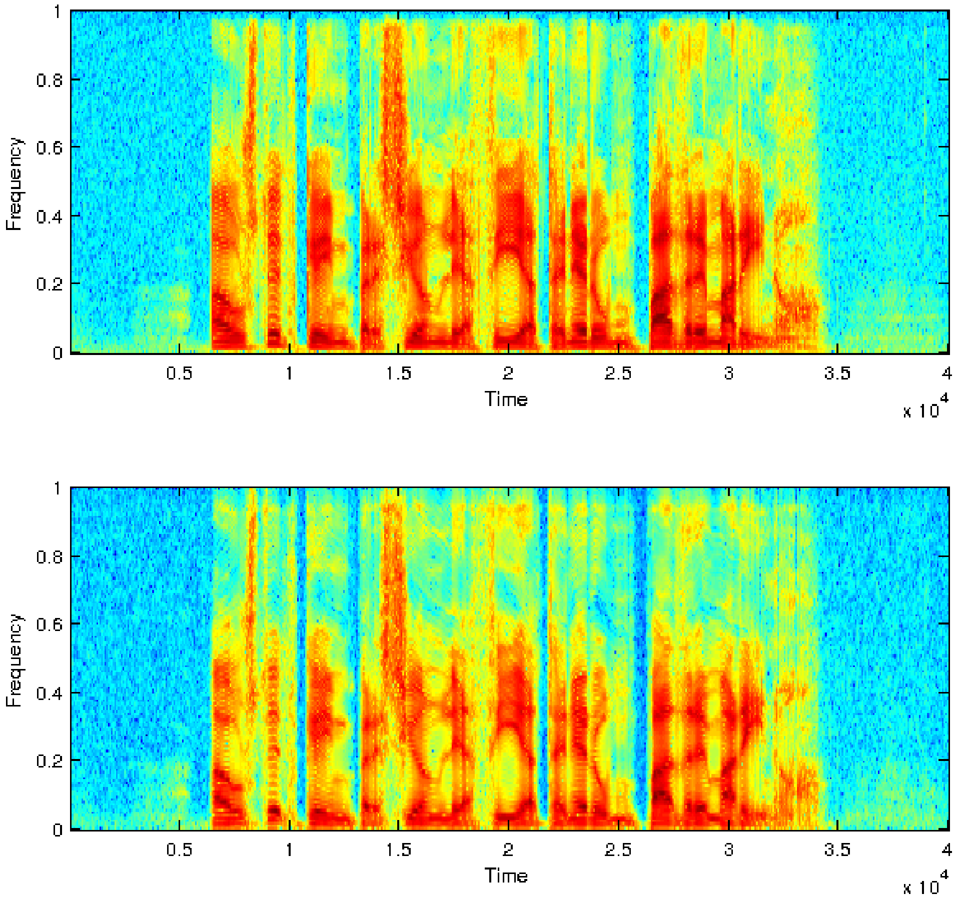

Figure 13: Spectrogram for the angry emotion of the original file (above) and the synthetic file after resynthesis (below)

- F0_MIN: Minimum fundamental frequency
- F0_MAX: Maximum fundamental frequency
- VOICING_THRESHOLD: Voicing threshold with respect to gain in the low-frequency band, so the speech frames under this value will be classified as unvoiced
- ZCR_THRESHOLD: Zero-crossings threshold. Speech segments that have more zero crossings than the threshold value are classified as unvoiced

So for the male voice the first two values are the almost the same for all the emotions and it is the other two the ones who change in each emotion. For the female voice this values are totally different than for the male, for example the maximum f0 is bigger than in the case of the male voice. If all the f0 values obtained after the feature

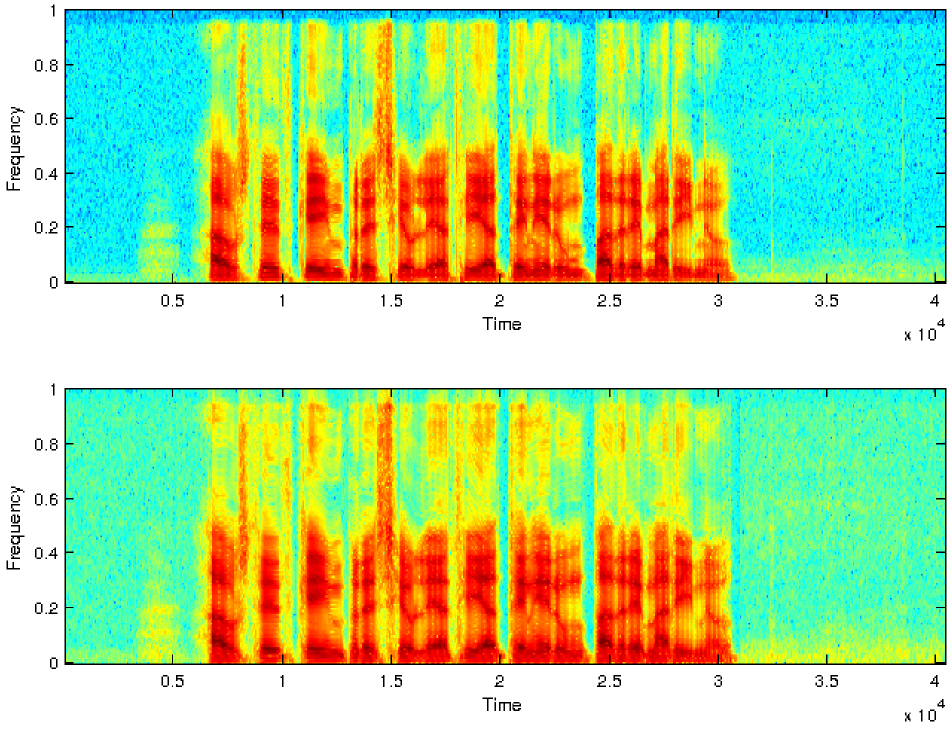

Figure 14: Spectrogram for the sadness emotion of the original file (above) and the synthetic file after resynthesis (below)

extraction (section 8.1.3) of all the files used for training are plot, the different f0 for each emotion can be seen in figure **??** in section A.4.

8.1.2 HTS Configuration File

In this configuration file the path where the features are going to be extracted is given, and also the streams that are going to be used in the training. The streams are like vectors where all the information of the features of the same type, that will be used in the training, are stored. In the experiment the next streams have been used:

- f0: fundamental frequency
- lsf1: spectral envelope LSFs
- gain1: gain
- flow: source LSFs
- hnr_i: harmonic to noise ratio with bands

Looking into this file or in the training script it can be seen that the dimension (or size) of this streams is 31 for the lsf (10 for lsf, 10 for the delta coefficients, 10 for the delta-delta coefficients and 1 for the gain), 10 for the flow, 5 for the hnr and 1 for f0.

8.1.3 Feature Extraction

The next step is the feature extraction of the audio files that are going to be used in the training. The features that are going to be extracted can be selected in the GlotHMM configuration file.

The streams that contains the features will be used in the training for building the voice model, so the features have to represent the voice. The extracted information for a file is stored in a binary file with cmp extension and is the one that will be used in the training step.

8.1.4 Training

Once the feature extraction is done the training step can be started. For this a folder with the features (cmp files) and a folder with the time alignment labels is needed.

The time alignment labels can be extracted using a front-end. For this a question file will be needed. The trained is HMM based with five states Gaussian and leaf nodes for the different trees (see section 5). For each training two models are generated due to a reclustering is applied to obtain better results. Once the training is done and the models are created, one thing that can be done is to realign the training labels using the model that has been created during the training step and train a new model with this realigned labels. This can be done as much times as wanted, and in the case of this project it has be done two times (so we have three rounds) with the male voice in both cases (dependent models y average voice) and with the female voice just with the dependent models due to that with the female average voice a lot of computation time is needed (several weeks).

For realigning the labels another HTK tool is used, in this case HSMMAlign see [?] and A.1. So in the end a lot of models are generated due to the reclustering and the realignment. For the dependent models 6 models are created, 2 for reclustering in each training, and 3 trains are performed. In the case of the male average voice the same 6 models are created but the adaptation is done for each emotion with the last one of the reclusterings models so in the end 3 models are generated for each emotion ,so 15 models for the male average voice.

As a test is going to be done some sentences for the test need to be synthesized (see section 9.1) and they have to be the best ones, so before the synthesis this sentences one of the generated models has to be chosen as the best one. This has been done as explained in section 9.1.

8.2 Adaptation

The adaptation consist in transfer the capabilities of one sepeaker to an average voice (in speaker adaptive training, sat) as is explained in 7. So basically the steps that has to be followed are the same that with the previous method (section 8.1), with the difference that this time an average voice model has been build with all the emotions to have a more robust model.

In order to do this all the extracted features (cmp files) for the previous method will be placed in the same folder, and the same goes for the time alignment labels, and the SAT (section 7.5) has to be set to one in the training script, so there will be only one model (the average).

So at this point is where the method differs of the previous one. Now is where the adaptation take place. So an adaptation to this model has been done with every emotion which generates a new model for each emotion. According with what is told in section 7 the type of adaptation is supervised and batch-mode, so different adaptation techniques could have been applied here like MLLR, CMLLR, MAP (see 7) or a combination of some of them like CMLLR + MAP, this is called CSMAPLR (also SMAPLR exist) [12]. The adaptation technique that has been used in this Thesis is the CSMAPLR adaptation. For the CMLLR 256 regression tree nodes have been used. Using MAP after CMLLR improve the average log probability per frame a little bit which leads in a better adaptation. Previously to this MAP technique two iterations of CMLLR have been done to get better results.

The CMLLR adaptation can be tuned a little bit with some thresholds which can change the depth of the adaptation, so the emotion level can be tuned with this threshold. Also changing the regression tree nodes can affect the adaptation. As the adaptation has been done using a good amount of data the regression trees can be big and a better node will be selected.

For the adaptation all the data of the training for one emotion have been used to replicate the experiment that were done with STRAIGHT, but a big amount of data is not required for a good adaptation.

8.3 Synthesis

Once the model are created the process for synthesis is the same in both cases with a little exception when synthesizing labels that have not been seen during the training. In the case of the dependent models (section 8.1) the models has to be changed to know this new labels, in the case of the adaptation this labels are given when adapting so the new models created new them. For this change a HTK tool is used and it is called HHEd (**??**).

So when this is done the first step is to extract the features of the label that is going to be synthesized, as it was explained in 6.1 using the models created during the training. This extraction is done with another HTK Tool called HMGenS (**??**). This tool extracts the lsf, flow,logF0 and hnr of the label file.

The next step is to extract information of the extracted features to generate the F0, LSF, LSFsource, HNR and GAIN to use the synthesis tool of GlottHMM to generate

the audio file. When synthesizing the global variance (GV) can be used or not. It is supposed to help when the parameters are been generated from the label file with the over-smoothing......AAAAAAAAAAAAalgo por aqui pero no se si lo que escribo esta bien o no. In the case of experiments done sometimes it improved the quality of the generated speech, but in others it introduced some sounds (like whistles). So in some cases the GV was used and in others it was not used. In general with the dependent models it was nos used and with the average voice it was used, but not in all cases.
Also for the synthesis the HTS engine can be used but requires some transformations to the models.

9 Results

In this section the results for the experiments explained in section 8 will be showed. Different subjective test has been carried out for male and female voiced evaluating the two methods used for building the models (8.1, 8.2).

9.1 Training and Test Data

For the different methods and genres the amount of data have been different.

9.1.1 Male voice

For the different techniques the amount of data for the male voice is:

- Dependent model (section 8.1): the same amount of data has been used for each emotion. The total amount of data per emotion is 489 sentences (around one hour of recording speech)

- Adaptation model (section 8.2): in this case the amount of data used for the training is all the data of the dependent models, which is 2445 sentences, and then all the data for each model to perform the adaptation.

Before synthesizing the test labels one of the models created has to be chosen as the better, for that reason a validation test has been done with 15 sentences of other speaker. Once the model has been chosen twenty sentences from the Albayzin contest have been used for the subjective test.

9.1.2 Female voice

For the different techniques the amount of data for the female voice is:

- Dependent model (section 8.1): with the female voice the amount of data is not the same for each emotion, for the neutral emotion less sentences (504) are used than with the other emotions (around 605 sentences) .

- Adaptation model (section 8.2]): for the female voice, as it has less quality and with the average model a more robust model is wanted, a lot of data is used: all the data of the dependent models (2922 sentences) plus the data of other 8 speakers (2808 sentences) , which makes a total of 5730 sentences.

For the female voice no so many models has been created due to that in the adaptation the amount of data is too big and it takes more than a week to perform one training. In the dependent models the same models than in the male voice were obtained but the realignment didn't have good results in this case.The validation and test sentences are the same as for the male voice. All this information is compacted in table 1.

#sent Voice	anger	happiness	neutral	sadness	surprise	average	validation	test
Male	489	489	489	489	489	2445	15	20
Female	605	603	504	605	605	5730	15	20

Table 1: Number of utterances used in train, validation and test

9.2 Test

As the purpose of the test is to compare GlottHMM with STRAIGHT the test has been divided into two parts.
In the first one the quality and the naturalness of the vocoder is tested, so two audio files are showed (A and B) and the next questions are asked in the test:

- Choose the file that represent better the emotion (A or B)
- Choose the file that is more natural (A or B)
- Choose for both files the level of emotion (from poor to very high)
- Choose for both files the level of naturalness (from poor to very high)

In the second one the test is focused on the speaker voice, so it is asked to choose the file (A or B) with the voice more similar to the original speaker.
In the test the listeners are not going to listen to all the audio files, so the files showed in the test have been randomized using Latin square (??), so that way nobody has control over the test.
In order to obtain enough results, three test have been performed. One for each dependent model and other one for the adaptation models. The reason for doing the adaptation test together was the difficulty to find listeners for the test.

9.3 Male Voices Results

Here the results for the male voices and the two different methods will be showed.

9.3.1 Dependent Model Results

The results for the dependent model for the male voice are showed in figures 15, 16, 17 and 18. Puede que haga 2 si las mezclo pero una al lado de la otra no creo que quepan y como son distintas no tiene sentido puedo agruparlas ES y MOS...

Faltan resultados del parecido de la voz al original, si estan pero en formato tabla en la parte SIM y MOS es Mean Opinion Score

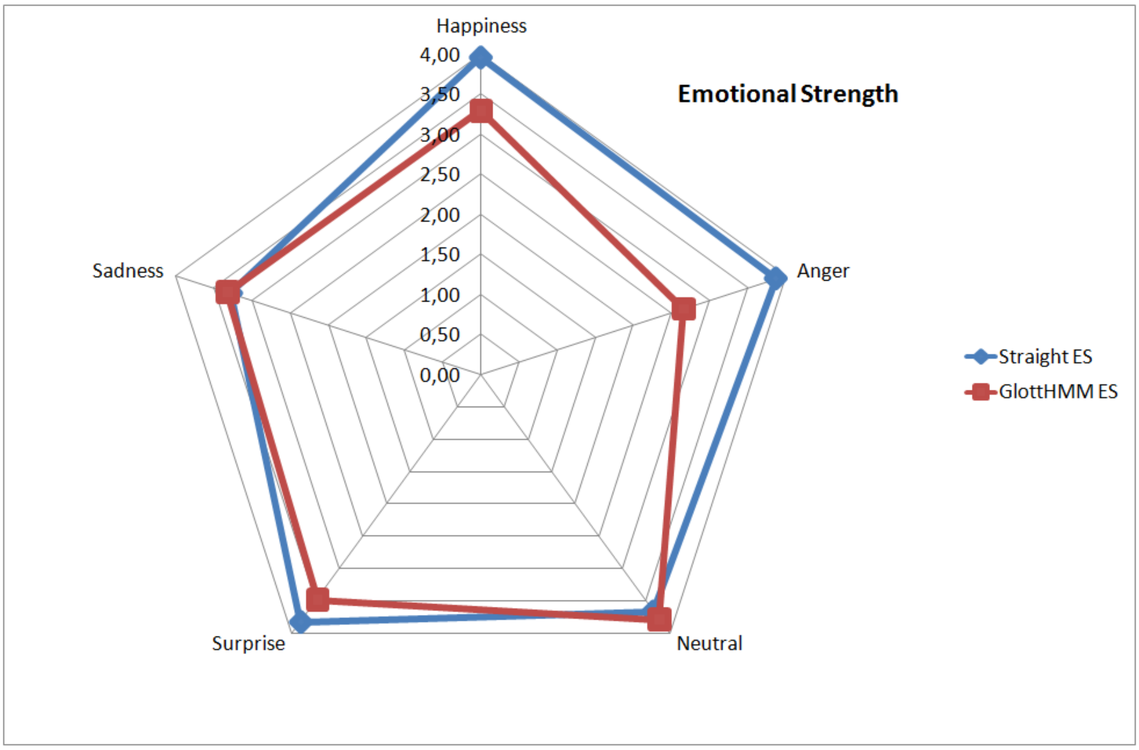

Figure 15: ES representation for the emotional strength

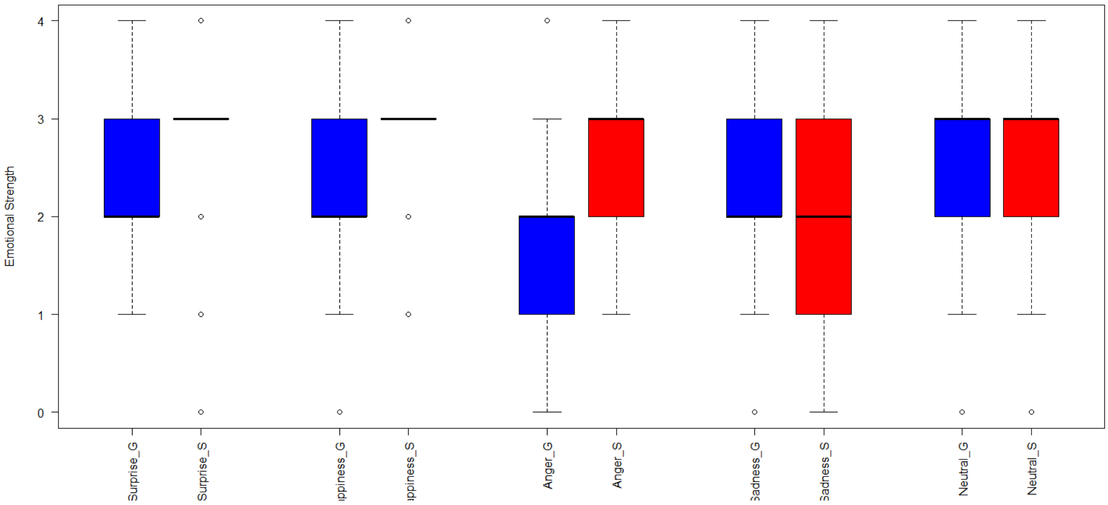

Figure 16: ES boxplot representation for the emotional strength

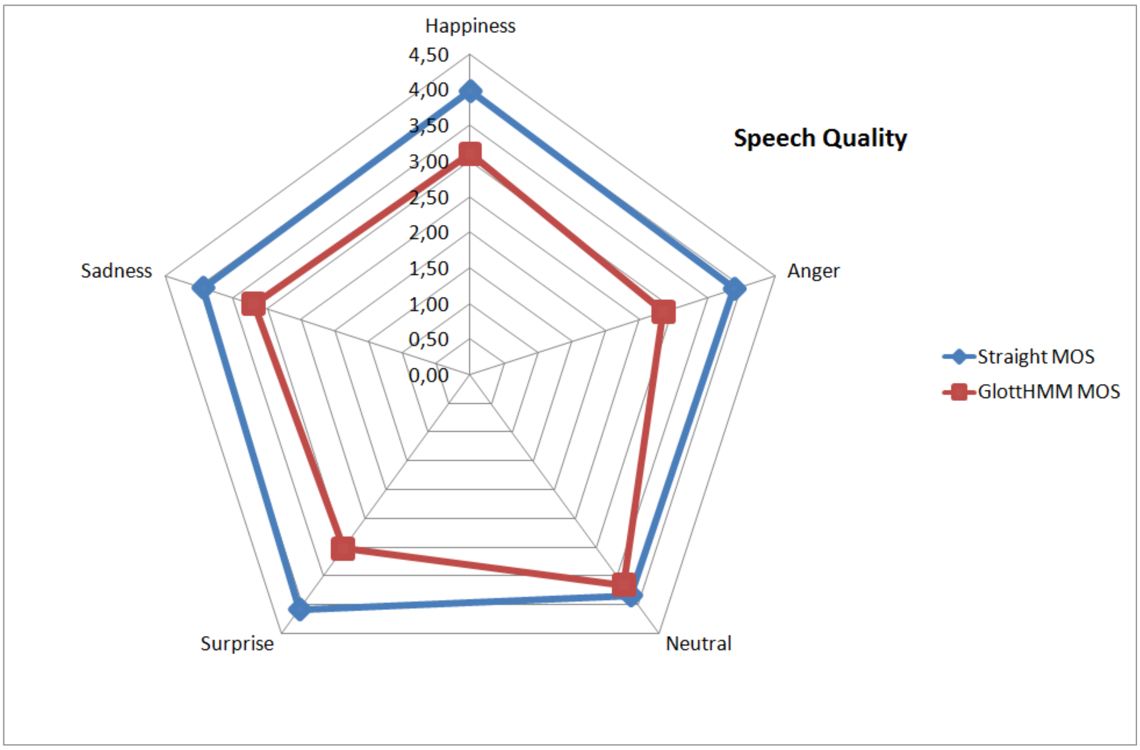

Figure 17: MOS representation for the speech quality

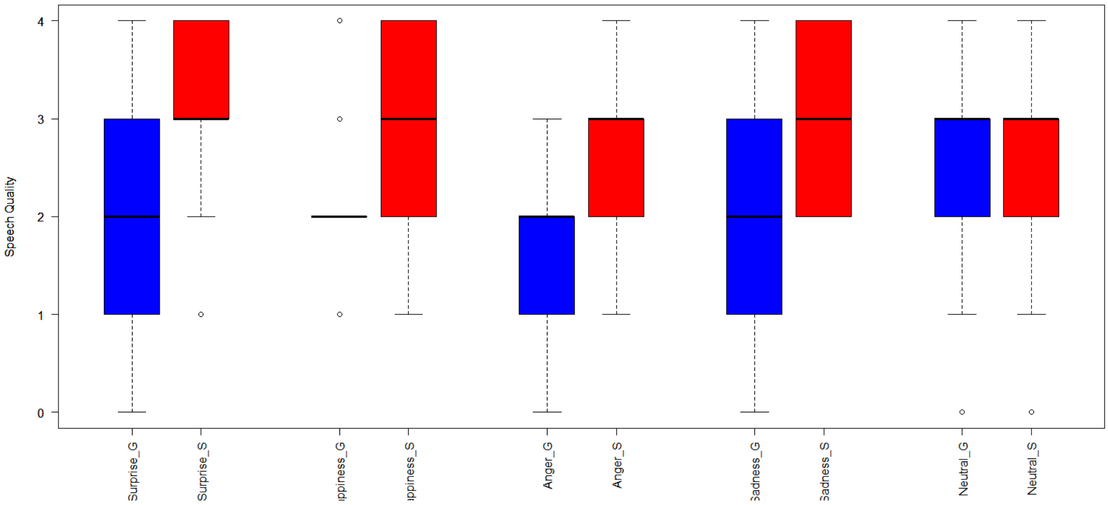

Figure 18: MOS boxplot representation for the speech quality

9.3.2 Adaptation Model Results

9.4 Female Voices Results

Here the results for the female voices and the two different methods will be showed.

9.4.1 Dependent Model Results

9.4.2 Adaptation Model Results

10 Conclusion

References

[1] M. University, "A brief historical introduction to speech synthesis: A macquarie perspective." last accesed 11-03-2014.

[2] J. Yamagishi, *An Introduction to HMM-Based Speech Synthesis.* 2006.

[3] K. Tokuda, H. Zen, and A. W. Black, "An hmm-based speech synthesis system applied to english," 2002.

[4] T. Raitio, "Hidden markov model based finnish text-to-speech system utilizing glottal inverse filtering," Master's thesis, Helsinki University of Technology, 2008.

[5] G. O. Hofer, "Emotional speech synthesis," Master's thesis, University of Edinburgh School of Informatics, 2004.

[6] J. Lorenzo-Trueba, R. Barra-Chicote, J. Yamagishi, O. Watts, and J. M. Montero, "Towards speaking style transplantation in speech synthesis," in *Proc. 8th ISCA Speech Synthesis Workshop*, 2013.

[7] M. Airaksinen, "Analysis/synthesis comparison of vocoders utilized in statistical parametric speech synthesis," Master's thesis, Aalto University School of Electrical Engineering, 2012.

[8] J. L. Flanagan, *Speech Analysis, Synthesis and Perception.* second ed., 1972.

[9] C. Darwin, *The expression of the emotions in man and animals.* New York D. Appleton and Co., 1897.

[10] D. Goleman, *Emotional Intelligence.* A Bantam book, Bantam Books, 2006.

[11] A. C. Aarón Blanco, "Inteligencia emocional." last accesed 11-03-2014.

[12] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorihms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, 2008. In print.

A Apendix A

A.1 HTK Tools

The HTK used is a patched version, so no all the htktools used are in explained in the HTKBook (see [?]) HSMMAlign no esta en htkbook HHEd HHEd basic operation is to load in a set of HMMs, apply sequence of edit operations and then output the transformed set. It is mainly used for applying tyings across selected HMM parameters. It also facilities for cloning HMMs, clustering states and editing HMM structures. For more information about this tool see [?]. HMGenS No esta en htkbook HERest Esta en el htk book

A.2 F0 examples

A.3 Latin Square

Esto de latin square mejor referenciarlo a algun sitio xd que llego bien de paginas yo creo

A.4 GlottHMM Configuration File Example