Expressive speech synthesis for a Radio DJ using Vocaloid and HMM's

Héctor Floría Rodríguez

MASTER THESIS UPF / 2013
Master in Sound and Music Computing

Master thesis supervisor:
Dr. Jordi Bonada
Co-supervisor: Marti Umbert
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

UNIVERSITAT POMPEU FABRA

Contents

List of Figures

List of Tables

Acknowledgments

During this Masters years, I had the honor to share knowledge with lots of people that has contributed to grow myself. I truly appreciate all the work done with and by all of them. Firstly, I sincerely thank Dr. Xavier Serra for giving me the opportunity to join the Music Technology Group. I would like to thank Dr. Jordi Bonada for supervising my master thesis and specially to the Phd candidate Marti Umbert for his strong support during the thesis development. Furthermore, my thanks are also to the classmates who have shared this journey to the Master's degree. Finally, I would like to thank my friends, APAC and family for their endless support.

Abstract

This master thesis deals with the modeling the speaking style of a Radio DJ speaker. In the context of the radio DJ speaker we designed a corpus that represents how radio DJs use to present songs being played on a radio show. A professional speaker has been recorded reading these radio shows into two different expressive styles. Style dependant models are trained with the recorded data. These can then be used to generate the expression parameters (syllable duration, pitch and dynamics) for any message to synthesize. The system is then modified by adding contextual information which is taken into account to train the models. The system is evaluatued to show whether this modification of the models helps to improve the speaking style models.

1 Introduction

1.1 Overview

The generation of speech with emotional expressivity and speaking styles variability is a current trend in speech synthesis research. The inclusion of expressivity to synthetic voices provides a better natural sounding synthesized voice. It helps to improve human-computer interaction and broadcast applications.

This thesis is related to the development of a virtual Radio DJ speaker. This application is based on broadcasted messages. These messages should convey to the listener the intended feelings and emotions of the radio speaker. As far as there is no feedback in this process, the expressive content is more important, it should stand on its own. The main objective of this project is focused on modeling the speaking style of a recorded speaker. With these style models, intermediate models would be created by model interpolation to synthesize utterances belonging to intermediate expressive Radio DJ speaking styles. These modeled speaking styles will be synthesized using state of the art synthesis techniques. To define our target speaking styles, we use the concept of arousal (calm, neutral or energetic) of the circumplex model of affect, explained more deeply in the following sections. Since a Radio DJ speaker uses neutral/positive mood there is no need to focus on the valence dimension.

The messages produced by the virtual radio speaker are based on real radio shows. These messages include a set of words/sentences that are typically used when a song is introduced, mainly regarding the song context (author name, band name, musicians names, album name, year, music style, etc.) and other aspects related to the radio show structure (relationship between sentences, sentence position, playlist flow, etc.). All this relevant information about a radio show has been identified, labeled and used to generate a set of 6 radio shows recorded. Each of these radio shows has been generated for a specific playlist of songs, representing in them the common aspects mentioned above.The recorded radio shows are the corpora that is used to train the speaking style models of the virtual radio DJ. The models are used to generate the expression parameters (syllable duration, pitch and dynamics) for any utterance (or message) to synthesize. These parameters are then used to control the Vocaloid synthesis engine to render the sentences.

1.2 Goals

Regarding to the global Radio DJ project context, the goals of this master thesis are:

- Analysis of radio shows to detect and analyze the keywords and common sentence structures, in order create basic radio show texts.
- Creation of a corpus-specific voice synthesizer in Spanish.
- Creation of the speaker models using the HTS (HMMs) framework.
- Evaluation of the HTS framework as a speaking styles modeler.
- Enhancement of the model through the addition of the labels (keywords) to the context.
- Evaluation of the model with this new context information.
- Adaptation of the HTS output parameters to Vocaloid synthesizer.

Other goals:

- Corpus recording.
- Automatic generation of radio shows through model sentences selection of the corpus.

The proof of concept presented in this introductory chapter can be summarized in the following figure. Figure 1.1 represents the main blocks of the Radio DJ Project which are studied developed in this MSc.

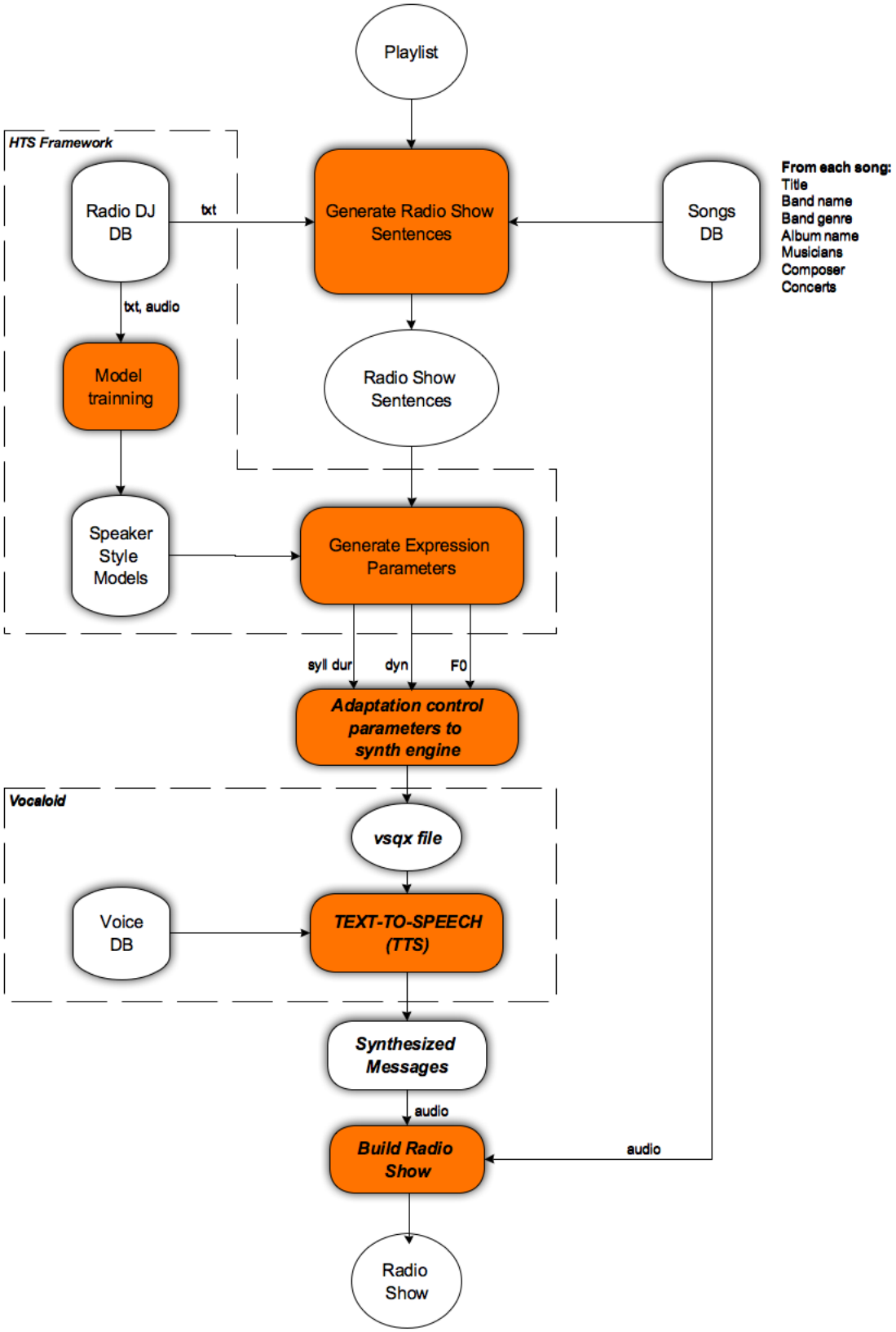


Figure 1.1: Thesis main blocks

2 State of the art

2.1 Historical background

Artificial human speech studies started centuries ago [HDSS06]. Some legends tell us about talking heads from the IX century. Further works appeared before electronic speech synthesizers, such as [DJ1791] and the the first Vocoder done by Bell Labs in the 30's. To better understand how the actual systems are working and how they have been reached this point and historical reading would be interesting. For a detailed description of speech synthesis development and history see for example Klatt [KL87], Schroeder [SCHR93], and Flanagan [FL72] [FL73].

2.2 Synthesis methods

Different synthesis approaches have been developed and used for speech according to Springer [SP08] and Taylor [TA09]. This approaches are synthesis techniques based on vocal tract models, synthesis by concatenation and signal processing modification, statistical parametric speech synthesis and unit selection synthesis.

As mentioned in [ZE09], the use of different methods and techniques has been related to the development of the computer resources. The creation of synthetic voices has progressed from knowledge-base to data-based one. Some evolution examples can be seen in [MO90] and [HU96].

The state-of-the-art text-to-speech (TTS) systems are mainly based on unit selection [HU96] or/and statistical parametric models [YO99]. The predominant approach to speech synthesis have been the unit selection techniques, according to [ZE09].

The **Unit selection method** is settled in the possibility to synthesize new sounding utterances using appropriate natural speech sub-words from a database. This means that the quality of the output depends on the quality of the recordings in the database. Unit selection synthesis is generally found to produce speech that sounds more similar to the target speaker than other methods, such parametric speech synthesis methods [KAR08]. Even that, in case a sentence needs phonetic and prosodic content that is not well-represented in the sub-words database, the quality of the output can be deteriorated. This involves building large databases, a difficult and costly process [BL03].

There exist two basic techniques in unit selection synthesis [HU96] [BL97]. Other approaches regarding the optimal size of the units have been studied and proposed as

[BL97] [DO95] [LI06] [HI04] and [BL02]. Furthermore, unit selection variants are being used in other voice areas as the singing voice [BON07] [FRE12].

Over the last years, **statistical parametric speech synthesis** has increased its popularity among speech synthesis methods [YO99] [ZE07] [ZE09]. In contrast with unit selection methods, that uses natural speech sub-words from a database, statistical parametric speech synthesis extracts parametric representations of speech (vocal tract, vocal source, prosody) and models them using generative models. The speech waveform is generated/reconstructed from the parametric representations. As mentioned in [ZE09], the most widely used generative models is HMMs. This synthesis is called HMM-based speech synthesis [YO99]. Furthermore, as unit selection, HMMs have been used also in other voice areas such singing voice [SAI06].

Statistical parametric synthesis has some advantages over unit-selection synthesis, mainly related to the flexibility due to the statistical process. In [ZE09] are described extensively. They group it as transforming voice characteristics, speaking styles and emotions; coverage of the acoustic space; multilingual support; footprint, robustness and speech recognition technologies.

On the other hand, unit selection methods are able to provide better (natural) sounding voices. Although, the best examples of unit-selection synthesis are better than the best examples of statistical parametric synthesis, the Blizzard Challenges [BL05] [BE06] [FR07] [KAR08] and others [AN12] test results, show that the quality of statistical parametric synthesis has reached a competitive level. Zen denotes that the quality issue comes from the process of reconstruction of the speech from the modeled parameters, that is not still ideal [ZE09].

Hybrid approaches that could combine unit-selection methods and statistical parametric synthesis, as Zen suggest in [ZE09], would be able to fill the gap, using from each method the best part of his performance. Several hybrid approaches have been developed. In [OK06] [AY 08] natural and generated segments are mixed. In [PL98] and [WO00] unit selection segment sequences are smoothed using statistical models and in [TA06] a unifying approach of unit selection and statistical parametric synthesis has been investigated.

The approach that will we taken in this master thesis is related with the target prediction hybrid approaches. As in [KA04] [HI04] [RO05] [YA06] and [KR08], the concept of using the Fo values, durations and dynamics are generated from statistical parametric synthesis (HMMs models) and then used as unit-selection synthesis input parameters. In our case the spectrum parameters values are not used because we are using with the timbre of the unit-selection method. These methods take the best performing part of both methods, the modeling of the statistical parametric method and the voice quality of the unit-selection method.

As mentioned before and as we will see in the next section, HMM-based speech synthesis has a better performance in speaking style modeling, which is the main target of

this Msc. Furthermore, as we have access to a unit selection based software (Vocaloid), we will take advantage of it to get better natural sounding voices.

2.3 Expressive speech synthesis

It has been recognized that vocal expression is one of the most important carriers of affective signals. Two centuries ago, Darwin [DA1872] already noted the importance of the voice as an effective channel. As Hofer wrote in [HO04], other authors, in more recent studies, have been looking for the specific vocal patterns of certain emotions and the relationship between vocal expression and emotion.

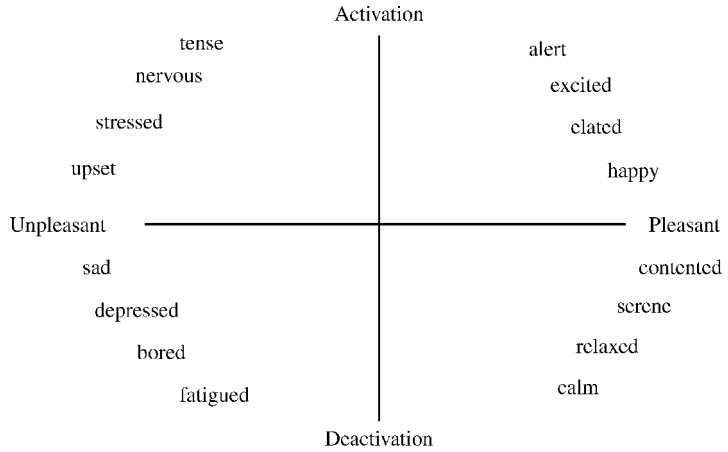
Attempts to add expression and emotions to the speech synthesis for achieving natural sounding speech is a long term goal of research in this field. Synthetic voices can synthesize good neutral read aloud voices in terms of intelligibility and naturalness [KI09]. These two properties are usually employed to describe the quality of synthetic speech. In several applications there is no need to apply complex expressive synthetic speech, due they just provide useful information to the user (eg: GPS systems). However, in the recent years there have been appearing other systems that bring human-computer interactions/conversations, such as virtual agents or virtual humans [TR08] [ROM10]. In these systems it is desired to make the speech synthesis more natural by adding more expressiveness to the speech.

In the context of this Msc, the RadioDj, although it is one way information source, adding expressivity for speaker characterization will give to the system more naturalness and reliability towards the listener.

2.3.1 Expressivity in speech

As mentioned, adding expressiveness to speech performance is a plus. In the case of this Msc, the focus is not in the modeling of emotional expressivity (such as sad, angry, funny), if not in the speaking style modeling (such as story-teller or discourse). Even though they are closely related. For that purpose, the mapping of expressivity into control parameters is important.

There exist works on emotions regarding singing voice [LB04] and musical performance [WG04]. In case of speech, a two-dimensional space (circumplex model of affect, shown in figure 2.1) was set in [RUS80] and [PRP05] relating arousal (activation) and valence (positive/negative). [SCH95] studied this model in speech and music.



Source: Feldman-Barrett and Russell (1998)

Figure 2.1: Circumplex model of affect

In the Radio DJ Project context, the circumplex model of affect will be reduced to 1-D, The different styles will be derived from the activation in the speech (calm to excited). Since we are dealing with the modeling of speaking styles and therefore only the arousal dimension will be taken into account. There is no need to have an "unpleasant" speaker style, so we might not to consider the use of the valence dimension. However, we have to take into account that the position in the valence dimension is not strictly in the middle (neutral) of the model. It should be considered moved towards the middle between neutral and pleasant.

2.3.2 Expressive speech synthesizers

Different approaches and techniques have been developed to generate expressive speech. This section includes a discussion on unit selection systems and HMM-based speech synthesis systems which are the techniques used in this Msc dissertation.

Other methods have been developed, as the ones presented in the Schröder review [SCHR01], such as formant synthesis (rule-based synthesis) and diphone concatenation (concatenative synthesis). For the singing voice, in [AL05] an expressive performance model focused on emotions for a singing voice synthesizer was developed using a ruled-base synthesis, based on the KTH synthesis of singing [KTH06].

Unit selection as an expressive speech synthesizer

As mentioned, concatenative methods such as unit-selection requires large speech databases. Regarding expressivity, the system should include a large database to generate expressive content for each type of emotion [BL03][PI06]. Other works had attempted to incorporate prosodic or phonologic strategies into unit selection [BA09].

HMM-based speech synthesis as an expressive speech synthesizer

The style modeling process of this thesis, will be done using the HMM-based speech synthesis method. There have been recent developments in HMM-based expressive speech synthesis that makes this speech synthesis technique an attractive and important way to follow and study.

The main problem with statistical parametric synthesis is that the spectra and prosody generated from HMMs tend to be over-smooth [BA09]. That impoverishes the quality of the final audio. However, it has other advantages in front of unit-selection methods that can provide very flexible speech modeling and generation.

In [BA09] these two state-of-the-art speech synthesis techniques have been applied to emotional speech, including perceptual test for six emotions (happiness, sadness, anger, surprise, fear and disgust). The results show that emotional speech generated from HMMs and from unit selection has similar speech quality, except from emotions having context-dependent prosodic patterns, where HMMs gives better results.

In [NO11] are described several developments of expressive speech synthesis about speaking styles, some of them used in this Msc. These techniques mainly focus on the reproduction and control of various speaking styles and emotional expressivity. There exist five related core techniques. Style modeling [YAM05] models and generates certain styles given sufficient amount of data. Style adaptation generates a target-style model by an adaptation of the neutral-speaker model trained with enough data, and uses it as an information acoustic model [TAC06]. Style interpolation [YO00][TAC05] can generate arbitrary intermediate expressions of multiple styles by interpolating between two style models. Style control [NO07] controls the intensity of style expressivity appearing in the synthetic speech. Style estimation [NO10] considers the inverse process of the style control to estimate the style intensity of the actual speech.

In this Msc, the **style modeling** and **style interpolation** will be studied and developed, following as a reference the Makoto's approach [TAC05]. The approach is based on a speaking style and emotional expression modeling technique for HMM-based speech synthesis, exactly the style modeling [YAM03][YAM05] and style interpolation [YO00] [TAC05] techniques mentioned before.

In style modeling the speaking styles and emotional expressions are statistically modeled. There exist two methods. Style-dependent modeling which each style is individually modeled and has his own acoustic model. Style-mixed modeling, where speaking styles and emotional expressions are treated as contextual factors; phonetic, linguistic factors and all styles are simultaneously modeled with a single acoustic model. As mention in [TAC05], both methods have almost the same performance, the approach used is the style-dependent modeling. It avoids the retraining of the models if new styles are added or removed.

In style interpolation, intermediate style models are obtained by interpolating the corresponding Gaussian distributions obtained from style modeling of each style. The interpo-

lation methods are the ones described in [YO00]. The model interpolation is applied to the spectral, F0 and duration parts. Furthermore, a style interpolation technique is proposed in [TAC05] to change smoothly from one style to another.

Focusing in this Msc, two styles will be modeled (calm and enthusiastic), from the recorded corpus-specific of the radio DJ. The modeling part will focus also on the inclusion of tags, in order to expand the context. One of the main goals is to see if the addition of important words of the corpus (band name, song, etc) as an information context will be useful to enhance the performance of the modeling.

2.4 Related Work

Within the frame of the Radio DJ project, previous work has been done in [UM10], where the objective was the generation of an expressive virtual radio DJ in term of emotions focusing on the transformation of a reference or anchor sentences instead of using a speech synthesizer.

Regarding applications related with broadcasted messages, in [FRA05] a system oriented to a weather forecast application is presented. More focused on speaking styles, in [TH06] the authors designed and implemented a set of prosodic rules for converting neutral speech, as produced by a text-to-speech system, into storytelling speech, for storyteller applications.

3 Design, labeling and recording the RadioDJ Corpus

This section summarizes the work done concerning the task of preparing the Radio DJ Shows to generate the style-dependent corpus database in Spanish.

3.1 Overview

A first task in the project is to create a corpus of typical sentences Radio DJs and group it to perform radio shows. In order to have a realistic and variate set of sentences and structures we have focused in real radio stations. This radio shows have been analyzed in order to obtain the relevant information given by the radio speaker. To build our radio shows, this information has been labeled with a set of tags to identify the content of it. The radio shows have been recorded with different speaking styles to evaluate the Framework as an expressive modeler.

3.2 Design of the radio shows

As far as the corpus is in Spanish the radio station that have been taken into account to extract the information are from Spanish radios such:

- Radio 3
- Flaix FM

Especially programs with one speaker where there are not dialogues and all the information given by the speaker is focused to the listener.

3.2.1 Analyzing radio shows, info collection and labeling

The analysis of the radio shows has been focus in order to extract:

- Common sentences & structures (Show introduction, Show end, songs presentation, etc).
- Relevant information related to the songs (Song name, song composer, band style, band name, etc).

3.2.1.1 Common sentences & structures

The analysis of the shows has ended in a set of rules in order to create sentence coherent structure able to create common sentences. Here we can see the organization strategy:

Sentence and context information labeling

For a clear sentence recognition, in terms of category, position possibilities, etc, a label has been created. The label consists in a four following sub-labels, that specifies how each sentence can be connected to other sentences in order to create coherent radio shows:

[CATEGORY BEFORE AFTER INFORMATIVE]

- **CATEGORY** denotes which is the type of the sentence regarding the position in the radio show workflow.

  Beginning sentence (S of start)

  Ending sentence (E)

  Pre-Song (B of before)

  Post-Song (A of after)

  Informative (I)

- **BEFORE** denotes if this sentence allows a sentences before itself.

  True if allows (T) but is not essential.

  True True if allows (TT) and it is completely necessary.

  False if not (F).

- **AFTER** denotes if this sentence allows a non-informative sentence after itself.

  True if allows (T) but is not essential.

  True True if allows (TT) and it is completely necessary.

  False if not (F).

- **INFORMATIVE** denotes if it is possible to add an informative sentence after itself.

  True if allows (T).

  False if not (F).

- NOTE: in case the labels AFTER and INFORMATIVE are both true, if needed, first will appear the informative label and then the next sentence.

Example:

- [S F T F] Bienvenidos a una nueva sesión de música con $Speaker.name, esto es $radio.name.

  S -> it is a Beginning sentence.

  F -> does not allow previous sentences.

  T -> allows sentences after itself.

  F -> does not allow informative sentences after itself.

This set of rules has been used to order and connect the set of common sentences extracted from the radio shows. Here we can see a sample of this sentences and his corresponding categories:

- Show Start:

  [S F TT F] Hola que tal! Como estás?,

  [S F T F] Bienvenido a tu radio!,

- Show End:

  [E T T F] Esto ha sido todo por hoy!

  [E T T F] Os dejamos con un tema de $band.name llamado $song.name para acabar.

- Pre-song sentences:

  [B T F F] Llegados desde $band.country, aquí tenemos a $band.name

  [B T T F] El tema se titula $song.name

- Post-song sentences:

  [A F T T] Acabamos de escuchar $song.name de $band.name

  [A F T T] Acaba de sonar $song.name, una canción de los $band.gentilic $band.name.

- Informative sentences:

  [I T T F] Vamos ahora con un poco de $band.style.

  [I T T F] Nos movemos ahora hacia el $band.style.

Relevant information related to the songs:

The song related information (context information) extracted from the shows has been labeled following five main groups of relationship: BAND, RADIO, SONG, ALBUM, OTHER. All the tags are listed in the Appendix A. As said before, the corpus database is in Spanish. Even that, we considered to include bands in the corpus with English names and also songs with English names. The problems derived from this decision will be mention later. Another thing to take into account is that the frequency of the labels have not been taken into account at the corpus creation time. Once the corpus has been created, they have been count in order to get the appearance frequency.

3.2.2 The generated Corpus

The corpus has been generated with the criteria of having more than 200 sentences or at list more than 10 minuts of audio. For achieving this values, has been necessary the creation of 6 text radio shows.

This radio shows have been created also taking into account all the possible sentences & structures, all the possible labels explained before. Trying to create meaningful sentences and structures with the least possible repetitions.

Also, the corpus has been done trying to give naturalness to the shows in terms of music coherence, that means thematic styles radio shows (Jazz, Metal, Funk, Mix,...).

The set of RadioShow generated can be seen in the appendix B.

3.2.3 Corpus Recording

The recording process has been done with a professional speaker and with the support of Marti Umbert. In order to test the HTS Framework as a expressive modeler the corpus (6 radioshows) has been recorded two time with different speaking styles:

- CALM
- EXPRESSIVE

This two styles has been chosen to have two different ways of speaking. Also a motivation is to give a way to a personalize the system. The calm style can be defined as a normal way to speak for the radio speaker (not neutral), with less variability in the voice. The expressive style can be defined as an enthusiastic way to speak with more variability in the voice and in the expressivity of the speaker.

The following table (cuadro 3.1) shows a resume of the recorded shows:

Shows characteristics
6 Radio Shows x Style
Total sentences: 259 (calm)/ 244 (expr)
2:00 ~ 2:40 each Show
Total Time: ~13 min x Style

Table 3.1: Shows characteristics

In the appendix B are listed the excerpt of Radio shows recorded.

Recording setup

Table following table summarizes the characteristic of the recording setup and figure 3.1 was taken during the recordings.

Feature	Description
Hardware	1x Neuman U87 microphone; 1 x cardioid (on-axis)
	1x mic stand, shock mount
	Anti-pop filter for cardioid mic
	Headphones for talkback
	Yamaha 03d mixing console
Software	Steinberg Nuendo
Script	Printed and placed in a music sheet stand

Table 3.2: Recording setup

Figure 3.1: Recording session

4 Methodology for expressive modeling

In order to model the expressive RadioDJ, since the beginning has been consider the use of the HTS Framework as a modeler. The HTS Framework has been used to create the speaker models. Moreover, to test the HTS as and expressive modeler, the radio shows have been modified in order to include in the context information labeled in the Corpus creation process.

4.1 HTS Framework

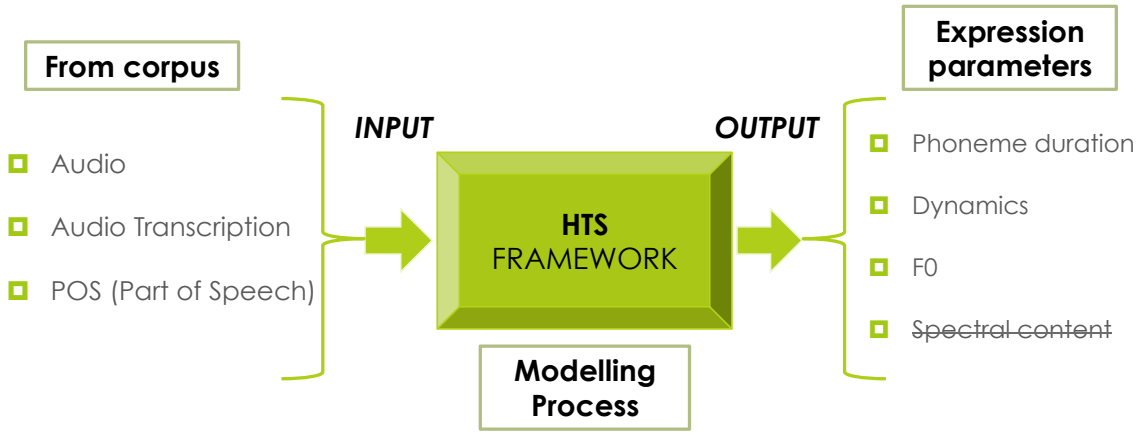


Figure 4.1: HTS Framework Diagram

The HTS framework (we can see a workflow in the figure 4.1) is able to generate a set of expression parameters for any message we would like to synthesize. That is achieved through the modeling process of a set of audio, transcription and label files from the recorded Radio Shows. This expression parameters are the ones that will control the inputs of the Speech Synthesizer (Vocaloid).

4.1.1 Input parameters

To perform the modeling process to obtain the speaker models and the corresponding expression parameters for any phrases to synthesize, the Framework needs a set of input data derived from the corpus (6 radio shows). This data corresponds to:

- Recorded audios segmented by phrases (.wav)
- Audio transcription of each (.trans)
- Label files of each phrase (.lab)

To obtain this set of files, the audio files and text files (radio shows) had been processed using in house algorithm and own codes.

Audio transcriptions (.trans)

The phonetic transcription has been obtained by means of the in-house transcription tool provided by Merlijn Blaauw which transcribes text into a sequence of phonemes using the symbols of the SAMPA dictionary, with accents and word and syllable boundaries.

Example:

/ [" d a . m o s] [" p a . s o] [a . o . r a] [" a l] [" x a s] [d e] [" x o n] [" k o l . t r e i n]
d a m o s p a s o a o r a a l x a s d e x o n k o l t r e i n

Figure 4.2: Transcription file

Label files (.lab)

The label file has been obtained through different processes (in house and open source tools). It contains information about each phoneme of the corresponding wav file, following the context-dependent label format for HMM-based speech synthesis.

p1^p2-p3+p4=p5 @p6-p7
/A:a1 a2 a3 /B:b1-b2-b3 @b4-b5 &b6-b7 #b8-b9 $b10-b11 !b12-b13 ;b14-b15 |b16 /C:c1+c2+c3
/D:d1_d2 /E:e1+e2 @e3+e4 &e5+e6 #e7+e8 /F: f1_f2
/G:g1_g2 /H:h1=h2 @h3=h4|h5 /I:i1_i2
/J: j1+j2-j3

Figure 4.3: Lab file format

The appendix C contains a description of all the parameters (figure 4.3) of the label file. And here we can se and excerpt containing the word “Coltrane” and in black the most important parts in order to understand the enhancement strategy:

o^n-k+o=l@2_0/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
n^k-o+l=t@1_1/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
k^o-l+t=r@0_2/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
o^l-t+r=e@4_0/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
l^t-r+e=i@3_1/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
t^r-e+i=n@2_2/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
r^e-i+n=Sil@1_3/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
e^i-n+Sil=xx@0_4/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1

Figure 4.4: Excerpt of label file

4.1.2 Segmentation, transcription and part-of-speech post-processing

Once we got all this data from the recorded audios, we performed some test with a set of models in order to have a first approach and also to check the correct performance of the framework. As mentioned, the generated audios this first models denoted that there were some problems with automatic segmentation and also some error in the phonetic transcription.

Phonetic Transcription

Some of this errors can be solved with a correct phonetic transcription of the english words. However in some cases the phonetic transcription does not exists. So that problem cannot be solve and we will drag this during the following processes.

E.g: Jazz -> Jas -> Xas

The phoneme J as it sound in the context of the word “Jazz” does not exist in the Spanish phonetics. The software assigns the current SAMPA symbol of the J like the word “Jarra”.

Audio Segmentation

In the case of the segmentation files, the problem can be divided in two parts:

- **English words misalignment**: In some cases were found misalignment in phrases cointainign english words.
- **Random misalignment**: Not all the problems came with the english words. Some time some words have been segmented well at the beginning and at the end, but in the middle the segmentation is not correct.

In the following images we can see two sentences and its corresponding segmentations, were are included the two examples explained before. The first .lab row corresponds to the pos-processed segmentation and the row below corresponds to the original one, we we can see the misalignments.

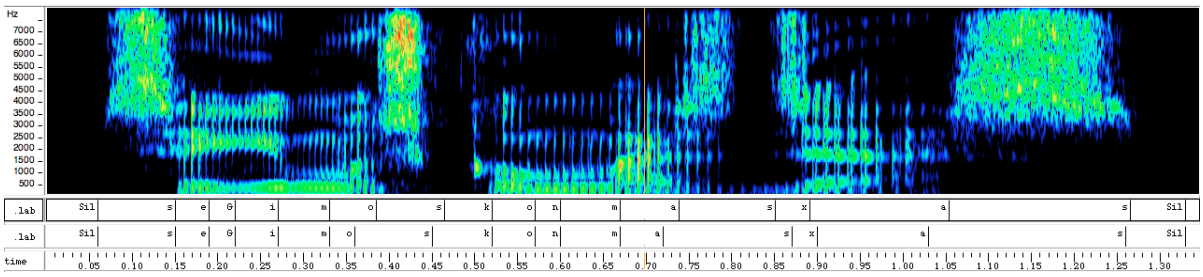


Figure 4.5: example 1 of segmentation correction

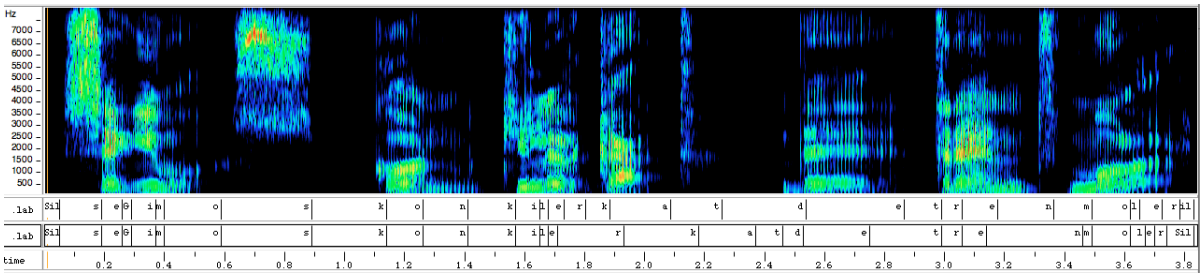


Figure 4.6: example 2 of segmentation correction

Part-of-speech post-processing

Due to the english words the POS, included in the .lab files explained before, had some errors. This errors have been corrected manually in order to obtain better models.

For example, the name "John Coltrane" has been automatically tag in the POS as:

- John: verb
- Coltrane: noun

That has been modified to:

- John: noun
- Coltrane: noun

In the figure 4.7 we can see the word coltrane tagged as a noun, but with the previous word tagged as a verb, tag "D:" (this word is John). Then in the figure 4.8 we can see that the "D:" tag has change correctly to noun.

o^n-k+o=l@2_0/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
n^k-o+l=t@1_1/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
k^o-l+t=r@0_2/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
o^l-t+r=e@4_0/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
l^t-r+e=i@3_1/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
t^r-e+i=n@2_2/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
r^e-i+n=Sil@1_3/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
e^i-n+Sil=xx@0_4/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**verb**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1

Figure 4.7: before post-processing

o^n-k+o=l@2_0/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
n^k-o+l=t@1_1/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
k^o-l+t=r@0_2/A:1_0_3/B:1-0-3@19-0&1-11#5-0$0-0!1-0;0-0|o/C:0+0+5/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
o^l-t+r=e@4_0/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
l^t-r+e=i@3_1/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
t^r-e+i=n@2_2/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
r^e-i+n=Sil@1_3/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1
e^i-n+Sil=xx@0_4/A:1_0_3/B:0-0-5@18-1&0-12#6-0$0-0!1-0;0-0|e/C:0+0+0/D:**noun**_1/E:**noun**+20@0+7&0+5#1+0/F:0_0/G:0_0/H:13=8@0=0|comma/I:0_0/J:13+8-1

Figure 4.8: After post-processing

Once all the files have been revised, the lack of quality produced by the set of segmentation, transcription and POS errors in the audios generated with the first models is now less noticeable.

4.1.3 Expression parameters

Once the modeling process by the HTS Framework is done, what we obtain with the speaker expressive modeling process is a set of expression parameters for any message we would like to synthesize.

This expression parameters are:

- Phoneme Duration
- Dynamics
- F0
- Spectral content (not used, it is used the Timbre from Vocaloid).

In the next figures can be seen this set of expression parameters (Spectral content, Dynamics, F0 and phoneme duration) of a calm audio file and expressive audio file. The differences between them are the ones that will be reflected in the models, the features that will characterize each model.

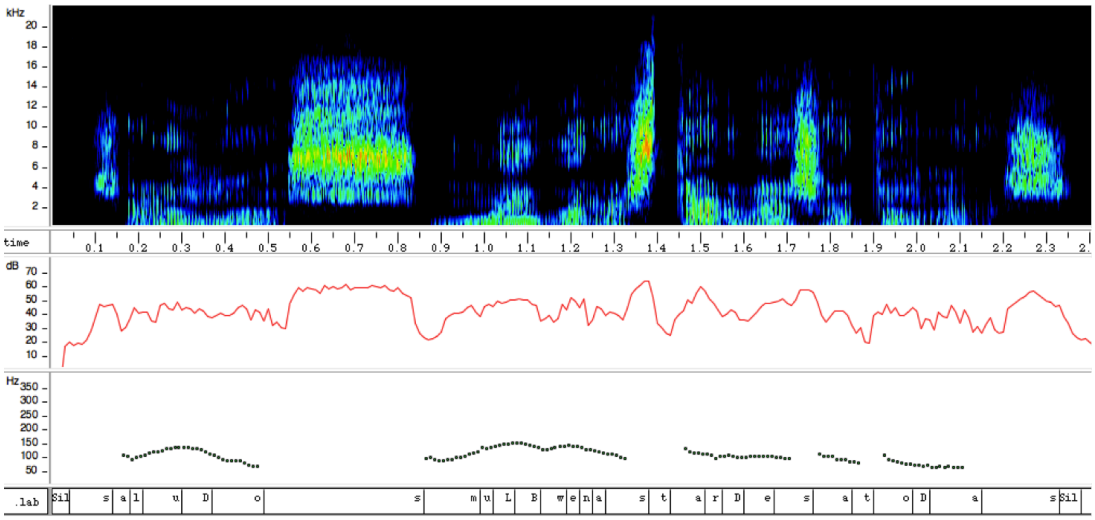


Figure 4.9: Expression parameters Calm audio

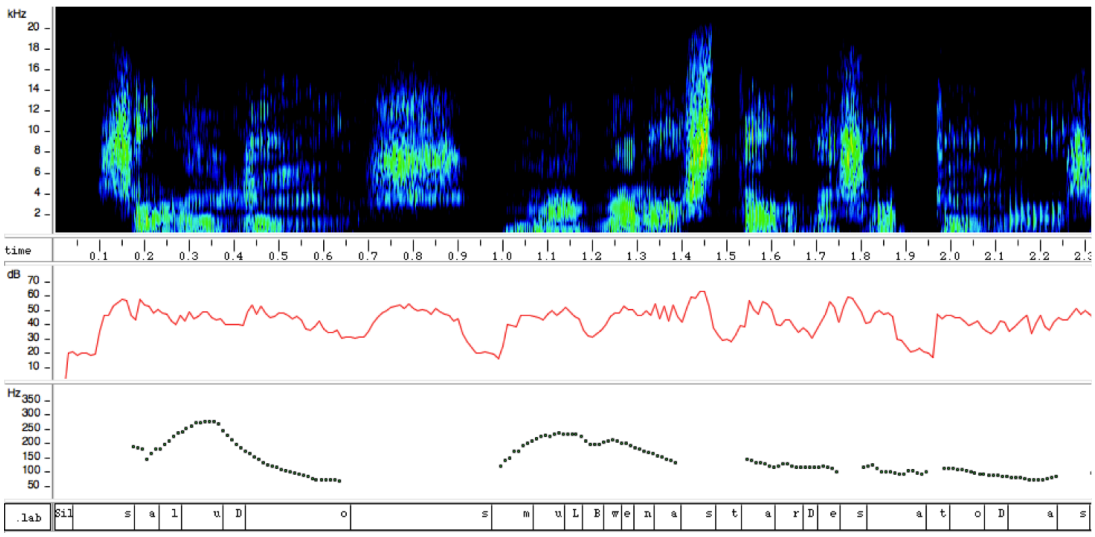


Figure 4.10: Expression parameters Expressive audio

4.2 Enhancement strategy

In order to test he HTS framework as an expressive modeler and to improve the base models (without context information) an enhancement strategy has been setup. It relies in the hypothesis that if the analyzed info contains context information about the shows, the framework will produce better expressive models. This context information is the one labeled in the corpus creation process, that refers to useful information about the songs, such as band name, song name, band style, etc.

The enhancement strategy consist in the replacement of the actual Part-of-Speech information (inside .lab file) by the tags of context information.

Example:

In the following example, the default POS info about the word “Coltrane” (tagged as a noun) has been replaced with the tags Bandname.

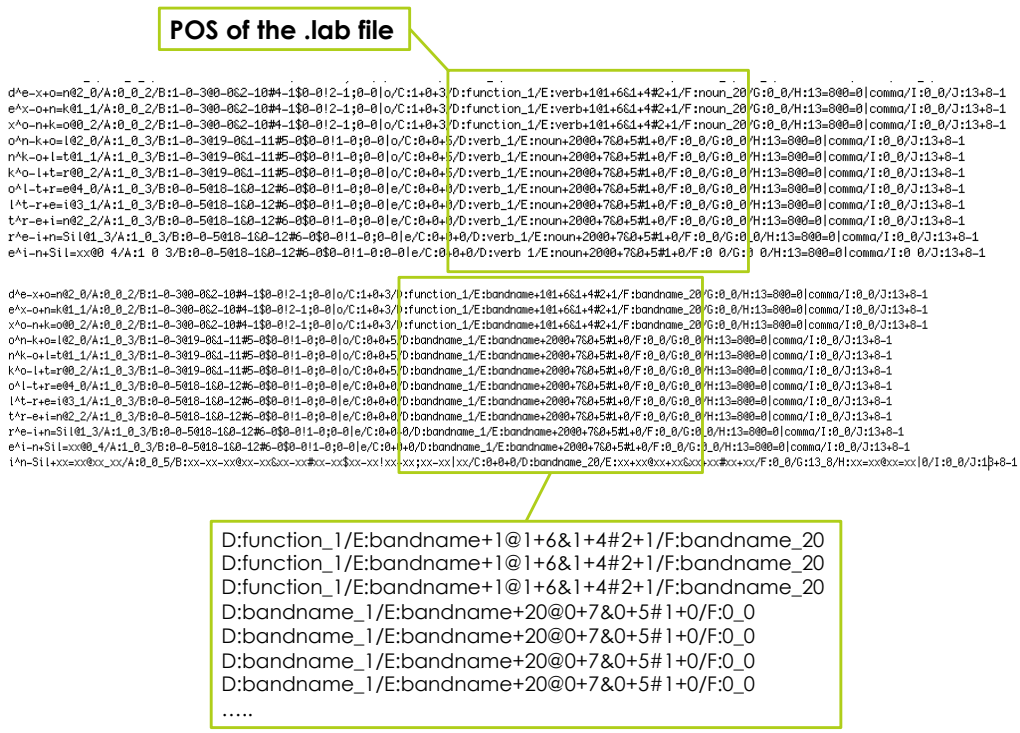


Figure 4.11: Example of the enhancement strategy

4.2.1 Datasets to perform the models

As mentioned, the enhancement strategy consists in the addition of the extra context information in the POS info. Before the addition of this new information, the default info given by the POS (.lab file) had a set of default tags with the following distribution:

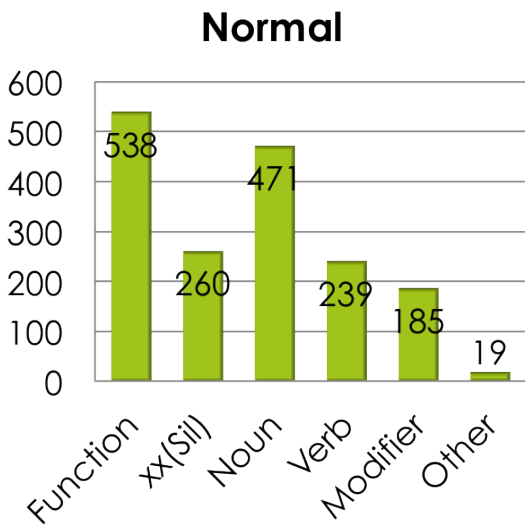


Figure 4.12: Normal Dataset

Once the enhancement strategy is performed and all the context information is added into the POS (.lab file), we get the following distribution of tags:

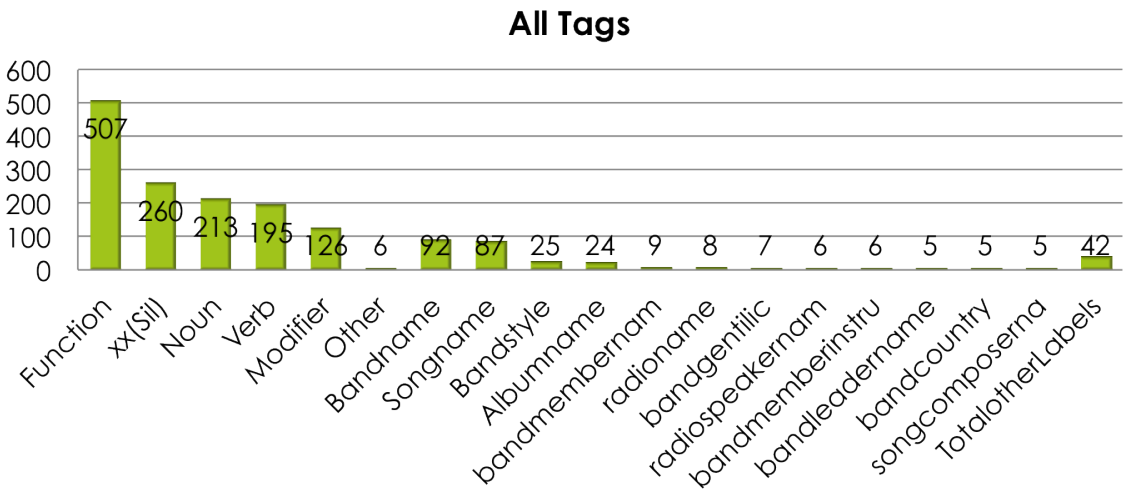


Figure 4.13: Distribution of all the context tags

We can see that once the strategy is applied the distribution of tags is not equal. There are a set of tags that are well represented through the Set of radio Shows (Bandname, Songname, Bandstyle, Albumname). This tags have been considered as the most important ones. Mainly, the rest of tags are not well represented because they have been thought as an extra information tags to create more dynamic the show in terms of given information.

From this global distribution of tags, another strategy has been setup. To solve that spread of info, we considered two approaches in order to create two reduced datasets with the context information available.

Reduced Dataset:

This dataset consists on keep the 4 most frequent tags and group the rest of tags inside main tags referred to given the root information. For example, the tags related to band info have been included in the main tag "BANDINFO".

'bandmembernameinstrument' 'bandmembername' 'bandleadername' 'bandgentilic' 'bandtourtime' 'bandleaderinstrument' 'bandleader' 'bandsytle' 'bandwebsite'

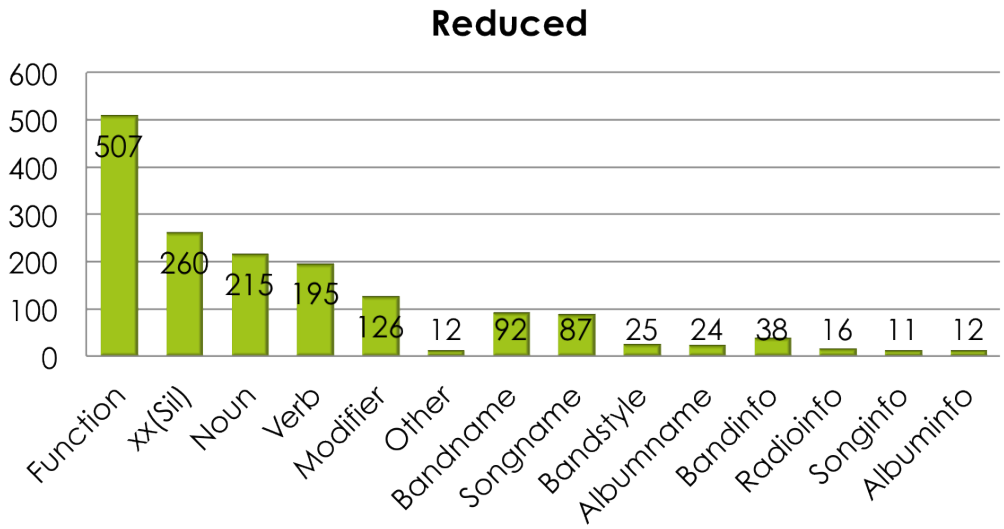


Figure 4.14: Reduced Dataset

4 Most Frequent Dataset:

This dataset consist in keeping only the 4 most frequent tags that are: 'songname': 87, ' ' bandname': 92, ' 'bandstyle': 25, ' 'albumname': 24, ' . This tags have been considered as the most important tags in the RadioDJ context.

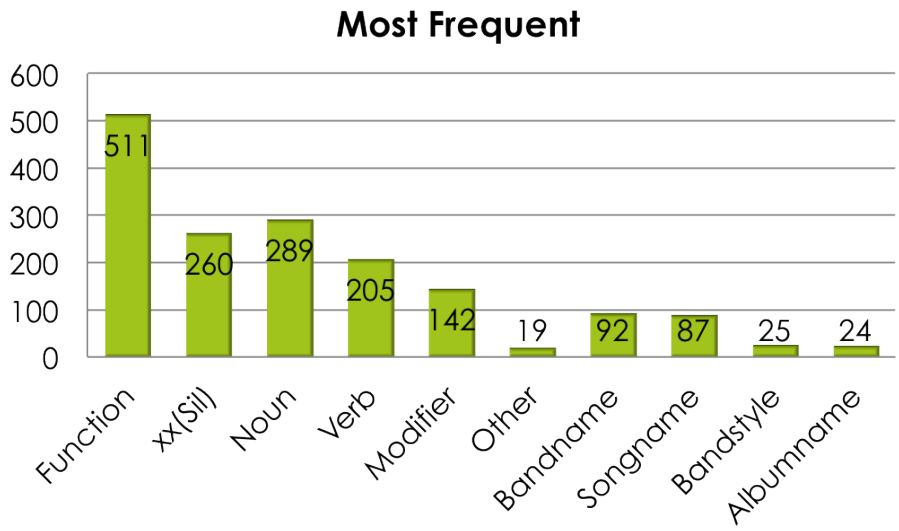


Figure 4.15: 4 Most frequent Dataset

Once we have applied this reduction what we get are 3 different set of data to test the framework:

- A normal dataset without context information.
- An enhanced dataset with 4 most frequent context information tags.
- An enhanced dataset with 4 most frequent context information tags plus a reduction of the rest of other tags.

Summary of datasets:

Finally, 3 datasets per style have been generated in order to test the HTS framework as an speech modeler and to test the hypothesis of the context information. What we would get is 3 speaker models per Style.

- Calm

Normal set, Reduced set, 4 Most Frequent set.

- Expressive

Normal set, Reduced setm 4 Most Frequent set.

4.3 HTS to Vocaloid Adaptation

The output parameters given by the HTS Framework (called previously Expression Parameters), are the set of parameters that are used to control the Voice Synthesis Software. The software used, as said before, is Vocaloid. The expression parameters given by the HTS Framework need an adaptation process because the input control parameters of Vocaloid Software are not the same. So, it the adaptation requires a certain degree of computation.
The set of expression parameters (outputs of the HTS) are:

- Phoneme duration
- Dynamics (Energy)
- F0
- Spectral Content (not used, because the Timbre of Vocaloid is the used one)

And the inputs of the Vocaloid are:

- Phoneme duration
- Dynamics (Volume)
- F0

4.3.1 Phoneme Duration Adaptation

The phoneme duration does not need any adaptation, it is directly mapped.

4.3.2 F0 adaptation

For the F0 values adaptation, the approach followed consist in the application of an interpolation to get samples different to 0. The F0 output of the HTS Framework, as we can see in the figures 4.9 and 4.10, has zero values in non-pitched areas (eg. consonants). As Vocaloid is a Singing Voice speech audio synthesizer it uses continues F0 and for each phoneme what we apply is a deviation to this continues F0 following the F0 input value. For that reason the 0 values have to be interpolated. To avoid abrupt changes a smoothing process with a 300ms window is applied.
This adaptation can be seen in the first graph of the figure 4.16 where the crosses corresponds to the original f0 curve, the red line corresponds to the interpolated curve and the green curve the smoothed curve in order to avoid the voice tremor.

4.3.3 Dynamics adaptation

The concept of dynamics in HTS comes from the Energy values. However, Vocaloid understands dynamics as volume parameter. The approach applied consists in detect the maximum value of the energy in the vowels. Then an interpolated curve of energy is created connecting this maximum values. This curve is normalized, between 0 and 1, with a center value of 0.7. Finally, we apply a smoothing process with a window of 150 msec. The process can be seen in the figure 4.16. In the 2nd row the interpolation between the first max pics of each vowel is done, and in the 3rd row the smoothing process and normalization processes are done.

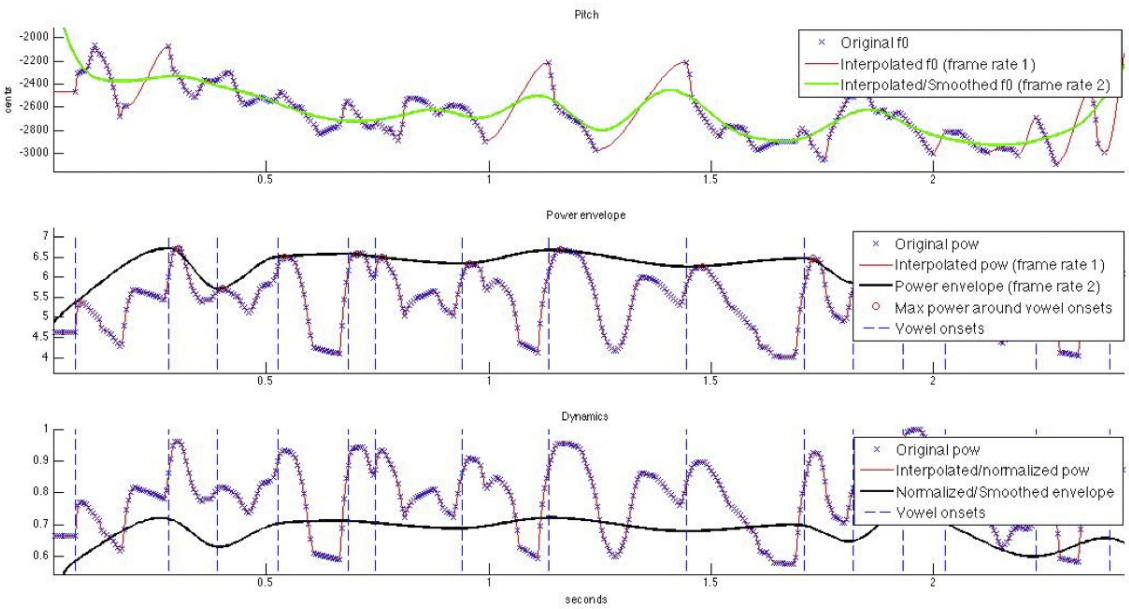


Figure 4.16: HTS to Vocaloid Adaptation

4.4 Orientative graphs

Before the Perceptual test, we had the possibility to compare the pitch contours and power values, used to synthesize the audios with Vocaloid. In the following figures per speaking style (calm and expressive) we can see the contour and power values used to synthesize the 10 sentences for the Perceptual test (chapter 5). In each figure are represented the values of the REC, NORM, FREQ and REDUCED models:

- The **REC** tag corresponds to the audios synthesized with Vocaloid controlled by the original input parameters extracted from the original audios (What is supposed to be the most accurate because used the original controls and not the modeled ones).
- The **NORM** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the Normal Dataset.

- The **REDUCED** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the Reduced Dataset.
- The **FREQ** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the 4 Most Frequent Dataset.

"Calm" results

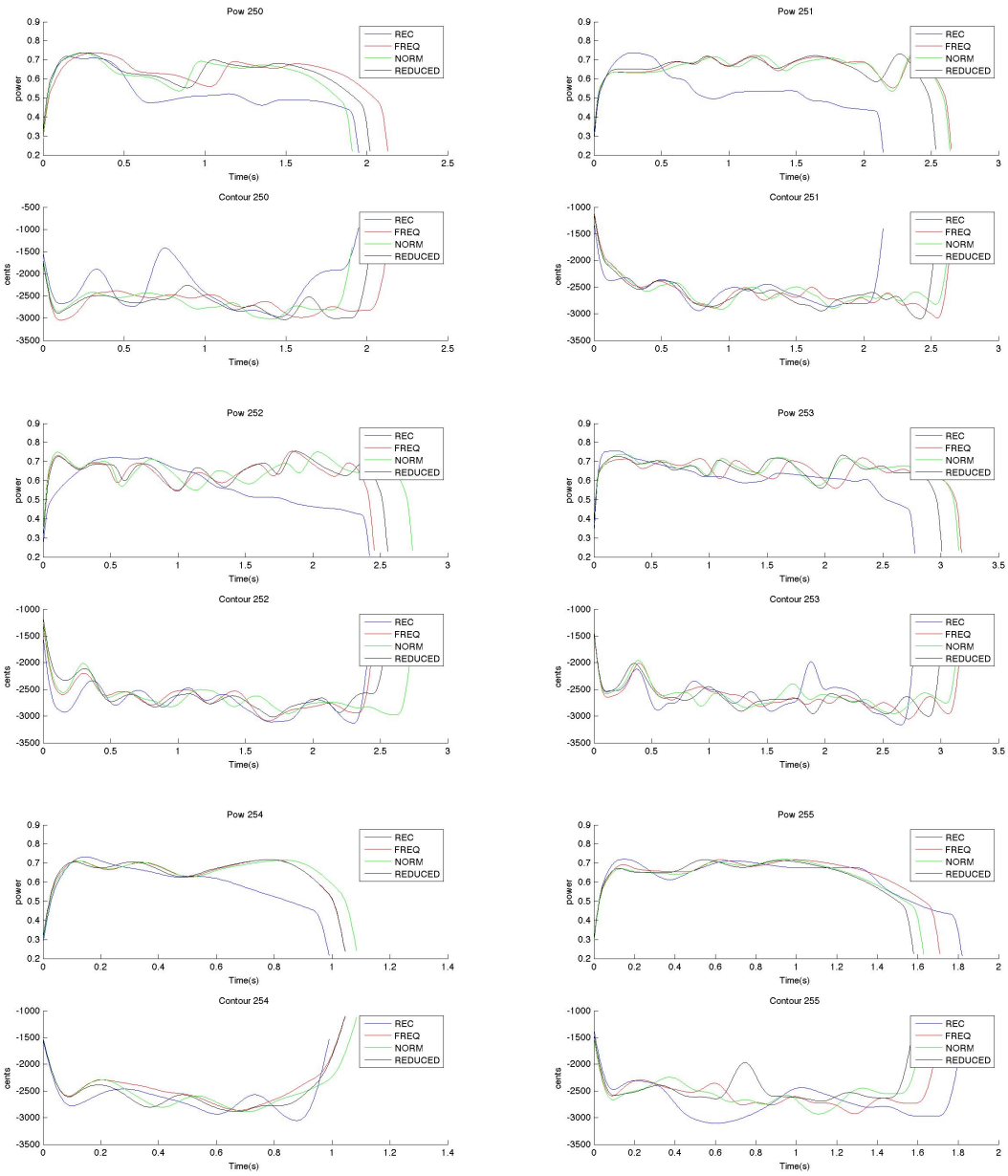


Figure 4.17: Pow and contour results

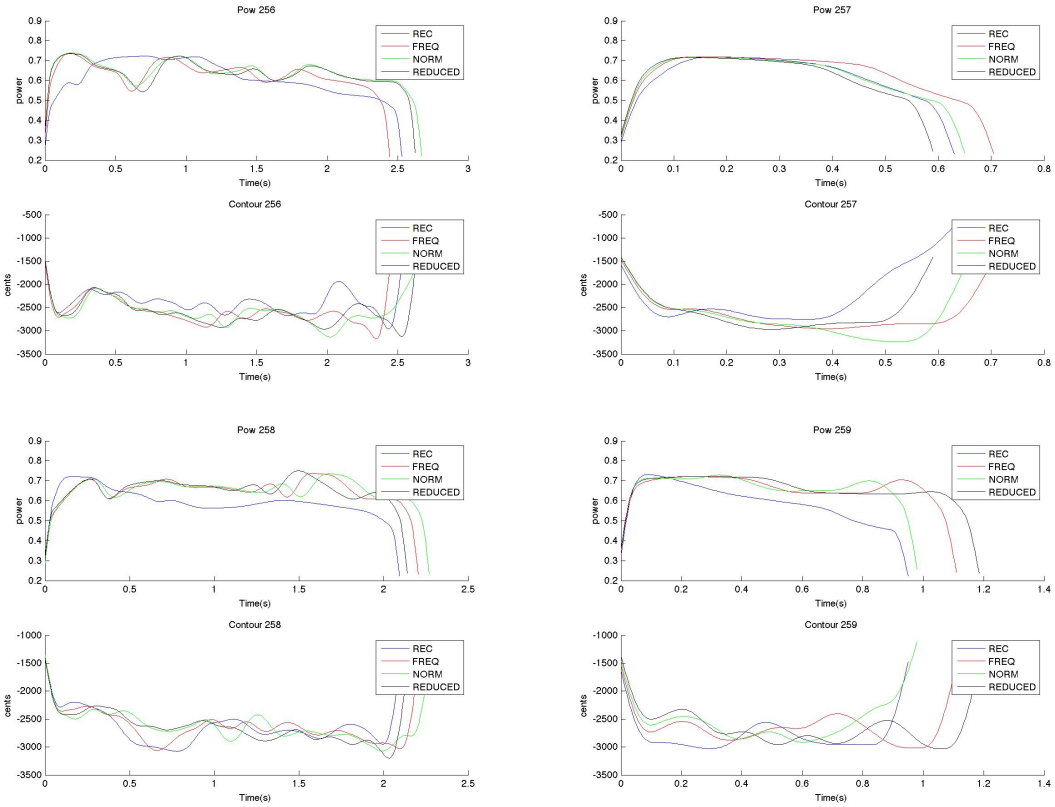


Figure 4.18: Pow and contour results (continuation)

“Expr” results

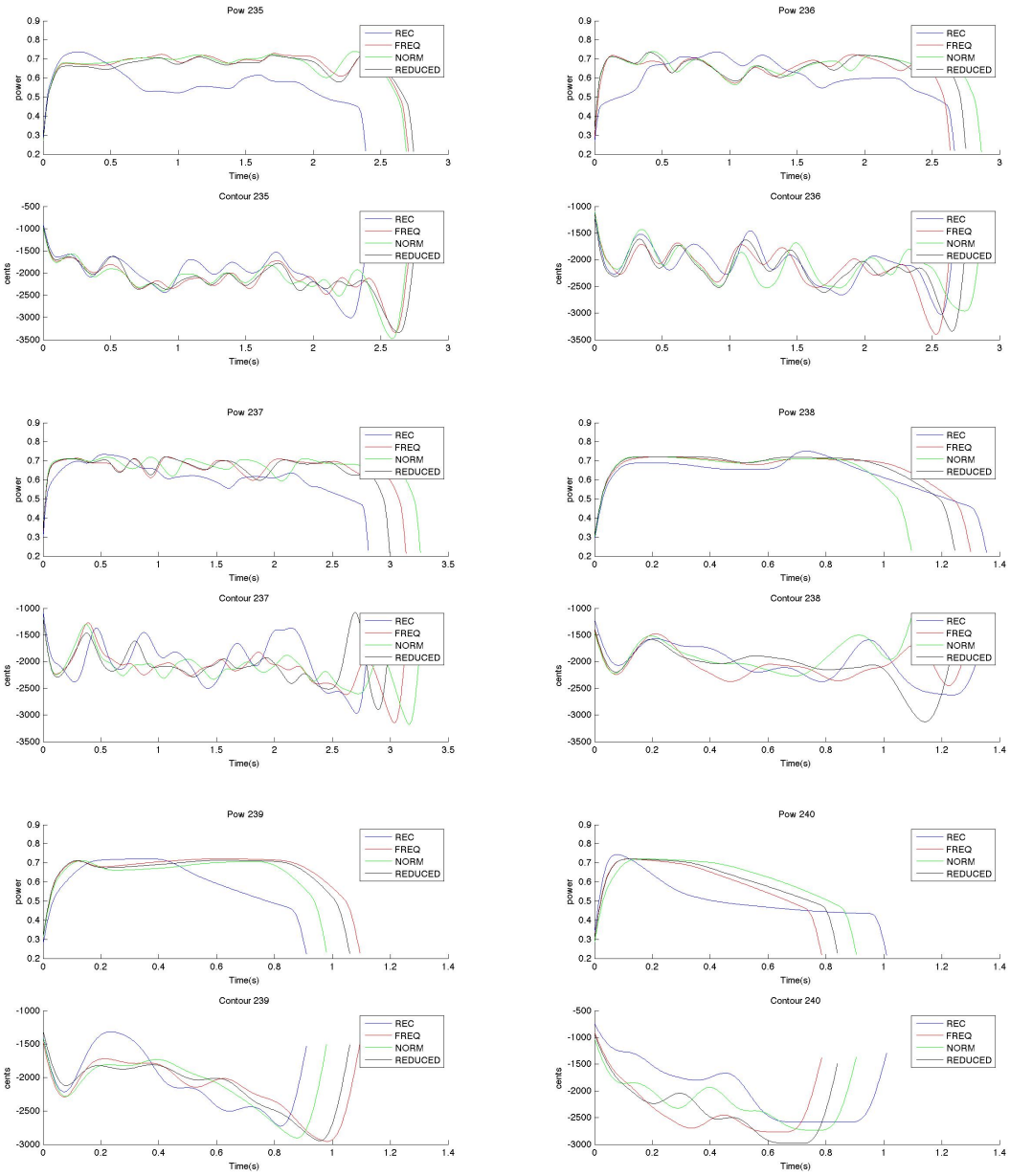


Figure 4.19: Pow and contour results

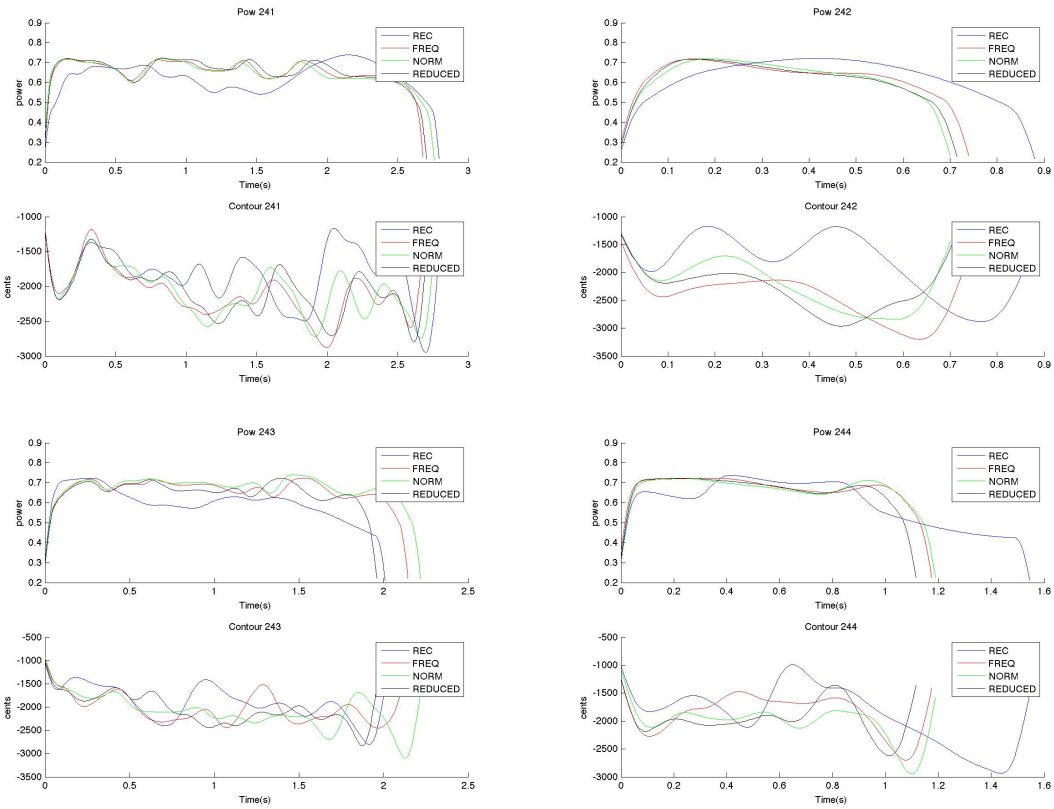


Figure 4.20: Pow and contour results (continuation)

Conclusions about the graphs

We can observe that, in calm and expr, the values produced with the models (Norm, Reduced, Freq) in the most of the cases follows the same shape. This shape is not far away from the REC one. The REC one is supposed to be the most real one because in uses the parameters extracted directly from the original audios. So, we could see from the graphics that the HTS is able to model the speaking styles, the most of the cases, accurately.

On the other hand, looking to the models values (Norm, Reduced, Freq), we can denote that even they have very close shapes, there are cases where the Reduced and Freq are different from the Norm one. These differences are due to the addition of the context information. E.g in calm 252 we can see differences in the power and the contours and in expr 239 and expr 240 the difference is more notable in the contours.

The conclusion extracted from the graphs will be verified or dismissed with the Perceptual test.

5 Evaluation and Results

5.1 Overview

This chapter focuses on the evaluation of the generated sentences and presents its results. A perceptual evaluation test has been carried out. More precisely, it has been evaluated which of the three generated models produces better results in terms of expressivity and naturalness, considering the RadioDJ frame. The test has been done to a group of 10 people non familiar with speech synthesis technologies, also most of them out of the engineering world.

5.2 Evaluation Methodology

5.2.1 Perceptual Test

The perceptual test has been divided in two parts:

- Part 1: asked the listeners to rate for each pair of audios presented: Which audio sounds more natural, in terms of expressivity.
- Part 2: asked the listeners to rate for each pair of audios presented: In which audio the specified words sound more natural (expressivity).

The answers have been obtained following a five-point scale for each pair of audios. The next table shows an example of a pair comparison between two models:

Audio A, Audio B
..... -2: definitely A
..... -1: Probably A
..... 0 : No difference
..... +1: Probably B
..... +2: Definitely B

Table 5.1: Pair audio comparison answer example

Each perceptual test part has been done with the two speaking styles Models (Calm and Expressive). And, as said, the comparison between models has been perform following a pair comparison test.

For each speaking style has been used the same comparison pairs between models (models listed in section 4.2.1), that are the following ones :

Part 1:

In part one, the user have been to answer six blocks of model-pairs. In each block two pairs of audios has been presented. In the following table we can see the combinations of models used.

Block 1	**Block 2**	**Block 3**	**Block 4**
2x Rec vs Norm	2x Rec vs Freq	2x Rec vs Reduced	2x Norm vs Freq

Block 5	**Block 6**
2x Norm vs Reduced	2x Freq vs Reduced

Table 5.2: Pair models comparisons of part 1

As a reminder, the **REC** tag corresponds to the audios synthesized with Vocaloid controlled by the original input parameters extracted from the original audios.
The **NORM** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the Normal Dataset.
The **REDUCED** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the Reduced Dataset.
The **FREQ** tag corresponds to the audios synthesized with Vocaloid controlled by the input parameters generated by the modeling process of the 4 Most Frequent Dataset.

Part 2:

In part two the REC model has not been considered because the addition of context information is only done in the generated models REDUCED and FREQ. It has no sense to compare the change of quality with the addition of the context information between non-generated and generated models. The user has to rate an specific word of the sentence presented, this word is related with the context information added to the models.

So, in the following table we can se the different blocks, models and words used in this part:

	Block1	**Block2**	**Block3**	**Block4**	**Block5**	**Block6**	**Block7**
Word 1	Maceo Parker	Maceo Parker	Pass the Piss	Pass the Piss	Funk	Funk	Omar Rodriguez
Word 2	Jizzy Free	Jizzy Free	Lets Go	Lets Go	<void>	<void>	Ernesto Soldevilla
Models	Norm vs Freq	Norm vs Reduced	Norm vs Freq	Norm vs Reduced	Norm vs Freq	Norm vs Reduced	Freq vs Reduced

Table 5.3: Pair models comparison of part 2

As a reminder, the models **Reduced** and **Frequent** are the ones that have been created with context information (section 4.2.1). The words listed in the table 5.3 have been tag during the context information process (section 3.2.1) as:

- Maceo Parker: band.name
- Jizzy Free: band name
- Pass the Piss: song name
- Lets Go: song name
- Funk: band style
- Omar Rodriguez: band info
- Ernesto Soldevilla: radio info

From Block 1 to 6 all the comparisons have been done between the Norm model (without context information) and the other two models. To see if the addition of this information enhance the models. For the Block 7, as the context information is the one added in the Reduced and not in te Requent. We decided to compare only the enhancement between them, and not to the Norm dataset.

The whole test has been done with **10** synthesized sentences per model.

The full Perceptual test can be seen in the Appendix D .

5.3 Test results

In the following section are presented the results of the perceptual test. The results are divided by model-pairs. That means that first for all we will see the results of the pair models that have the REC model plus other models, then the ones with the NORM models plus other models and finally the the model-pair of the enhanced models.

Furthermore, the results are presented in boxplots. This is the easiest way to visualize the results, since the shortest observation, the lower quartile (Q1), the median (Q2), the upper quartile (Q3), and the largest observation are shown in the boxplots. As a reminder, take a look to table 5.1, to check the possible test results.

5.3.1 REC Model vs NORM, FREQ and REDUCED models

As mentioned before, the REC audios are the ones synthesized with the original control of the original audios. Our hypothesis is that this should be the best model because in has original control paramenters and not the modeled ones with the HTS framework.

In the following boxplots we can see, per pair-model, the results for the both sentences presented for the CALM speaking style models:

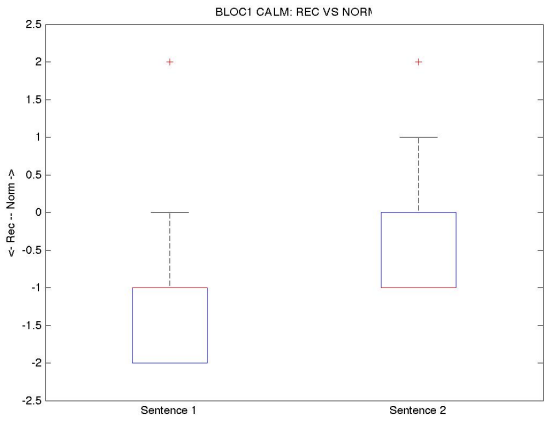


Figure 5.1: Bloc1 Calm: Rec vs Norm

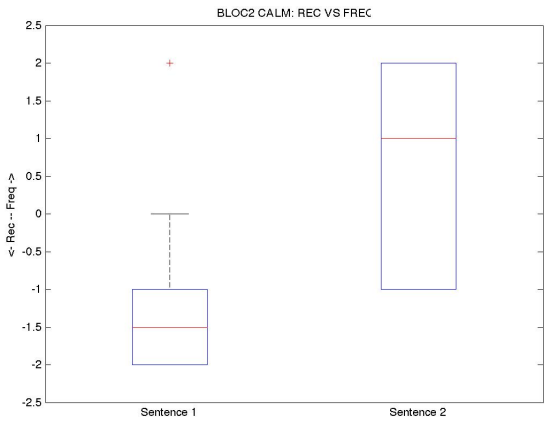


Figure 5.2: Bloc2 Calm: Rec vs Freq

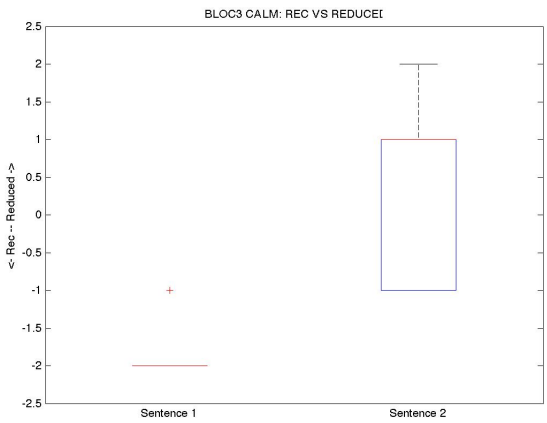


Figure 5.3: Bloc3 Calm: REC vs Reduced

The graphs show us that our hypothesis is accomplished in all the pair-model comparison of sentence 1. In the case of sentence 2, we can observe that when it is compared with the enhanced models, the results tend to be not as positive as in the sentence 1. Sentence 2 synthesized with the REC model has some problems in the words “Jizzy Free” and “Let’s Go”, where the prosody seems like a question. It might be the problem of this spread of opinions through the respondents.

The next three plots show the results for the same pair-models but with the EXPRESSIVE speaking style models:

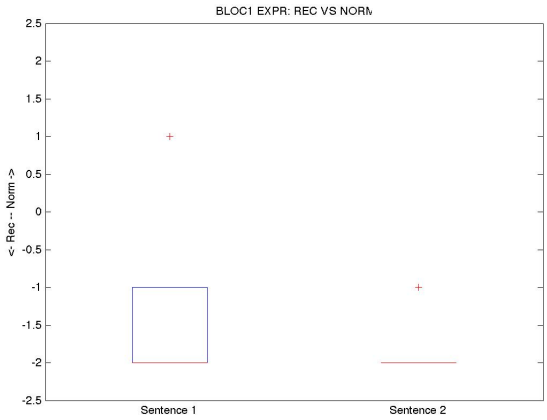


Figure 5.4: Bloc1 Expr: Rec vs Norm

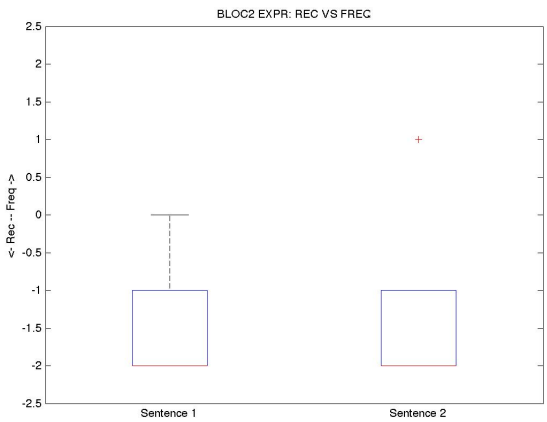


Figure 5.5: Bloc2 Expr: Rec vs Freq

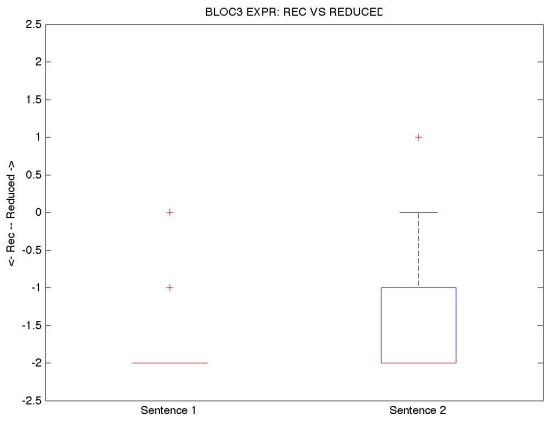


Figure 5.6: Bloc3 Expr: REC vs Reduced

In the case of the expressive speaking styles models, our hypothesis is also accomplish. In all the sentences the synthesized with the REC models have the best results.

5.3.2 NORM Model vs FREQ and REDUCED models

In the methodology we have set a enhancement strategy. It hypothesis relies on the addition of context information to the models in order to achieve a better modeling result. For that reason we wanted to compare the synthesized audios with the normal model (without enhancement) with the audios synthesized with the enhanced models (Freq and Reduced).

In the following boxplots we can see, per pair-model, the results for the both sentences presented for the CALM speaking style models:

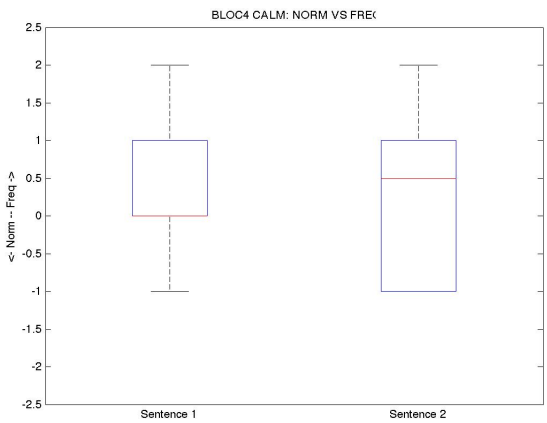


Figure 5.7: Bloc4 Calm: Norm vs Freq

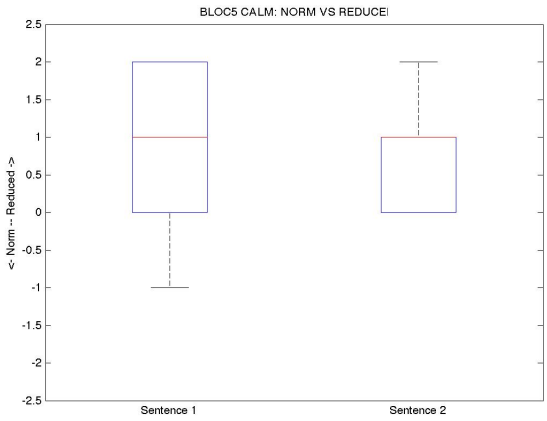


Figure 5.8: Bloc5 Calm: Rec vs Freq

As we can see, the results show us that the Freq and Reduced models tend to be prefered to the users than the Norm model. In any case, the users preferred the Norm model as definitely the best. Even that, some the number of people that considers the both audios equall is also important. So, we can consider that the addition of context information in the case of the Calm speaking style models is useful and our hypothesis seems to be successful.

The next three plots show the results for the same pair-models but with the EXPRESSIVE speaking style models:

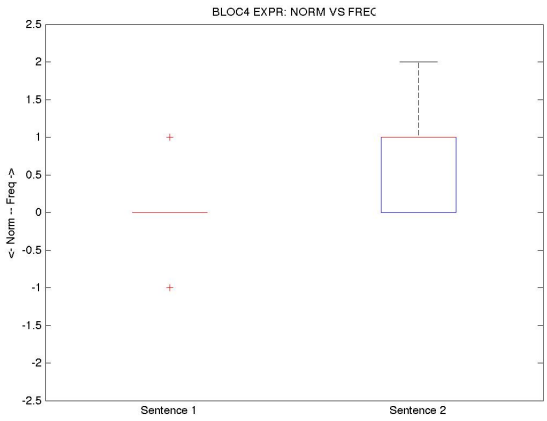


Figure 5.9: Bloc4 Expr: Norm vs Freq

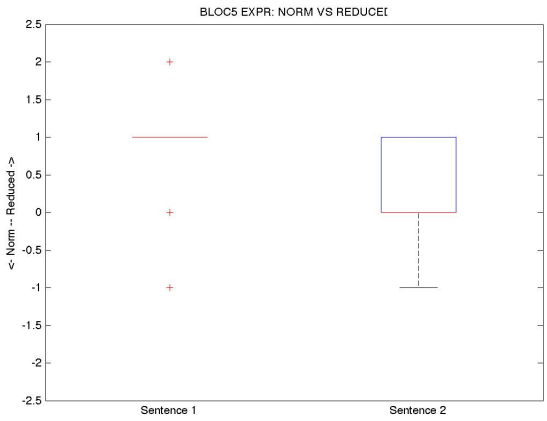


Figure 5.10: Bloc5 Expr: Norm vs Reduced

In the Expressive models the results also show a tendency of consider the enhanced models better than the normal one, but the number of neutral answer (considering the both model equal or non better one from other) is bigger.

5.3.3 FREQ model vs REDUCED model

As far as, we have perfomed approaches of the same enhancement strategy by creating two types of models with different clustering data criterias, we wanted to see if the differences between one and other.

The following graphs show the results obtained for the Calm and Expressive speaking styles:

Figure 5.11: Bloc6 Calm: Freq vs Reduced

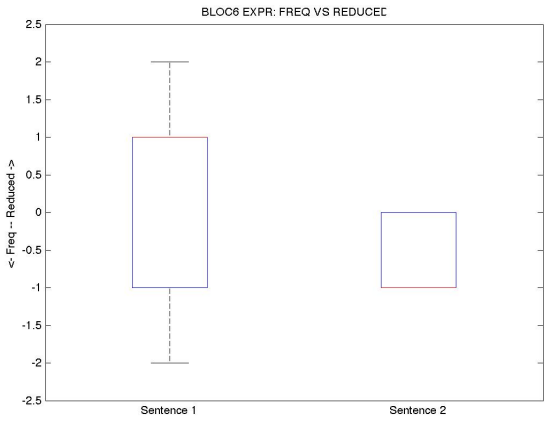


Figure 5.12: Bloc6 Expr: Freq vs Reduced

The results show that the tendency is to consider the REDUCED model better in almost all the cases. Even that, it is not totally clear because the number of neutral answers is also big.

Conclusion

As we have seen, in general the enhanced models (Freq and Reduced) perform better than the Norm model. But, we have been comparing the quality of the audios looking at the whole sentence. Right now, we cannot denote if this differences between the Norm models and the enhanced models is due to the specific tags (specific words) used in each approach or due to the way of grouping of the tags that modifies the whole sentences. For that reason we have perform the PART 2 of the perceptual test, asking to rank specific words of the sentences.

5.3.4 Results of the PART 2 of the Preceptual Test

As the enhancement strategy consist in add context information to the models using common meaning tags, we wanted to know if the models gives to this tagged words better expressivity. The test has been done comparing words of the that have been tagg in the enhanced models with the following taggs Band.name, Song.name, Band.style. Furthermore, we have ask to compare between the two enhanced models, one of the tags that the Reduced models has and the Freq model don't: BandInfo.

BandName results

Here we can see the results for the tag Band.name with the word "Maceo Parker" and "Jizzi Free" for the CALM speaking style models:

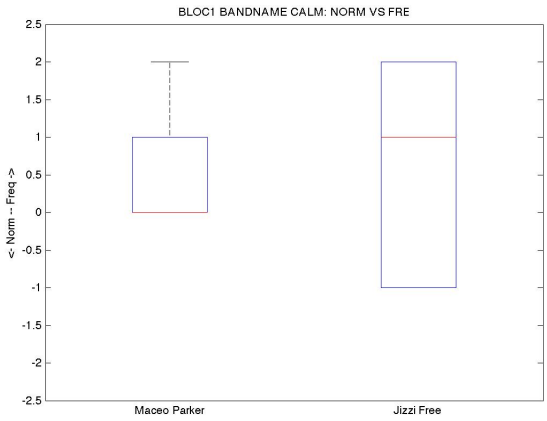


Figure 5.13: Bloc1 BandName Calm: Norm vs Freq

Figure 5.14: Bloc2 BandName Calm: Norm vs Reduced

We can observe that the models with the word tagged get better results. Even that, the neutral answer has also importance.

The following two plots shows the results for the EXPRESSIVE speaking style models:

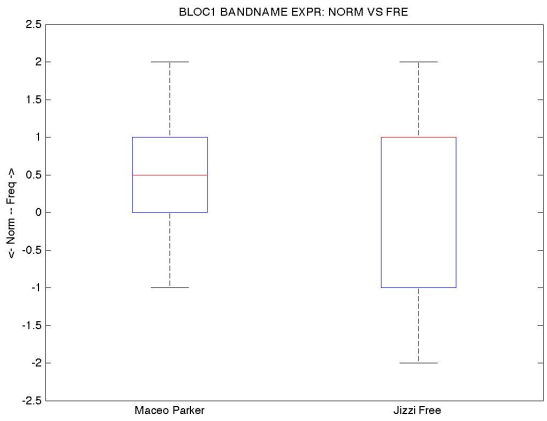


Figure 5.15: Bloc1 BandName Expr: Norm vs Freq

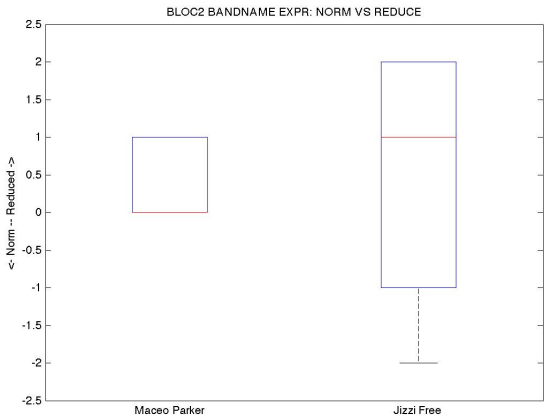


Figure 5.16: Bloc2 BandName Expr: Norm vs Reduced

In the case of the expressive models the the tagged words obtain better results but not as clearly as in the calm models.

SongName results

Here we can see the results for the tag Song.name with the word “Pass the Piss” and “Let’s Go” for the CALM speaking style models:

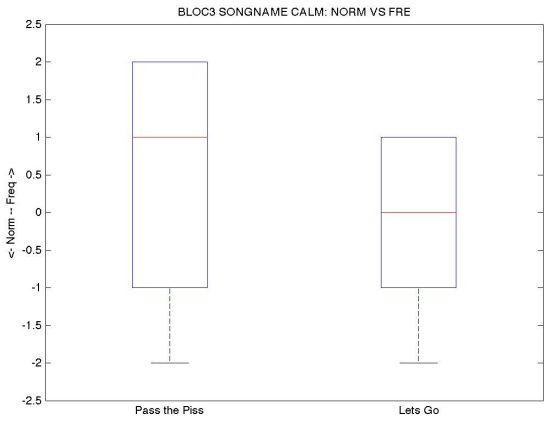


Figure 5.17: Bloc3 SongName Calm: Norm vs Freq

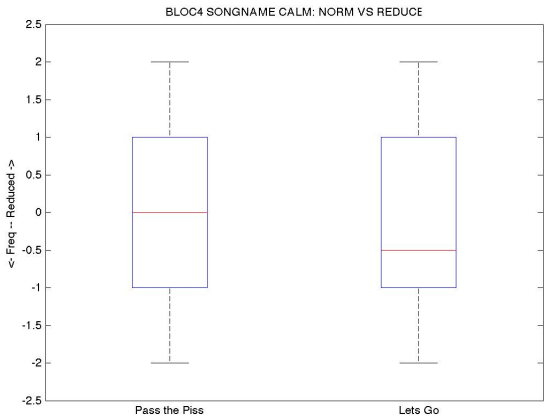


Figure 5.18: Bloc4 SongName Calm: Norm vs Reduced

In that case we cannot denote a preference for any of the three models compared. Even the Freq and the Reduced model don't have better significant results from the Norm model.

For the EXPRESSIVE peaking style models these are the obatined results:

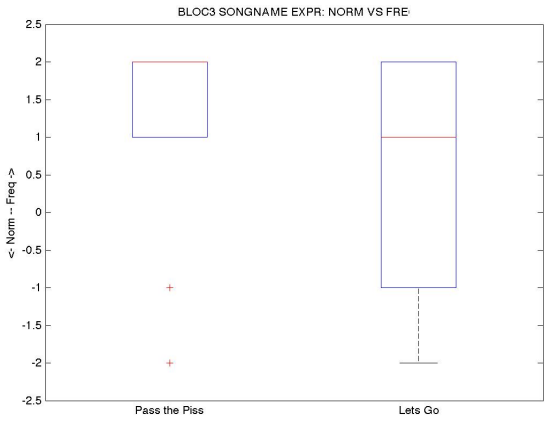


Figure 5.19: Bloc3 SongName Expr: Norm vs Freq

Figure 5.20: Bloc4 SongName Expr: Norm vs Reduced

The results show that the FREQ models is has modeled with more expressivity the tagged word but in the case of the Reduced model shows the contrary.

BandStyle results

Here we can see the results for the tag Band.style with the word “Funk”” for the CALM speaking style models:

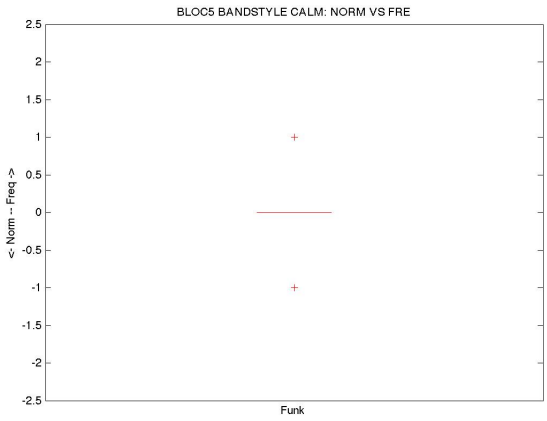


Figure 5.21: Bloc5 BandStyle Calm: Norm vs Freq

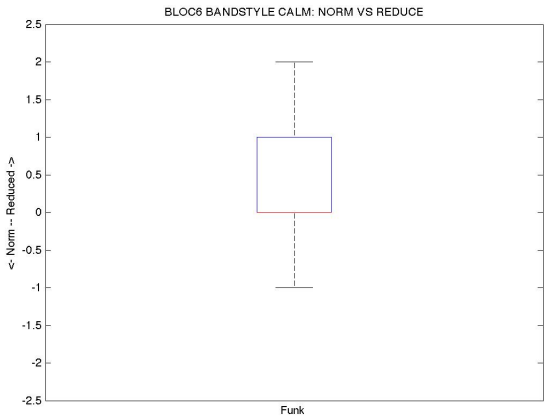


Figure 5.22: Bloc6 BandStyle Calm: Norm vs Reduced

The following two plots shows the results for the EXPRESSIVE speaking style models:

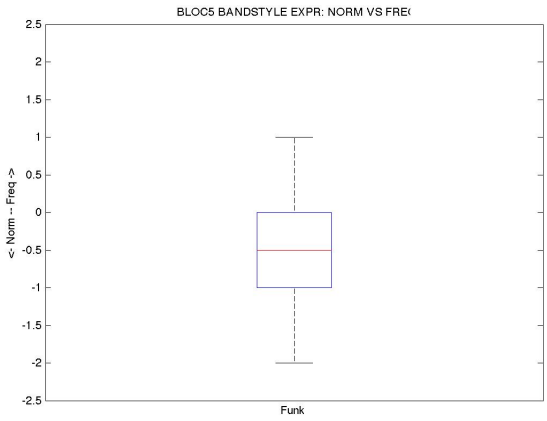


Figure 5.23: Bloc5 BandStyle Expr: Norm vs Freq

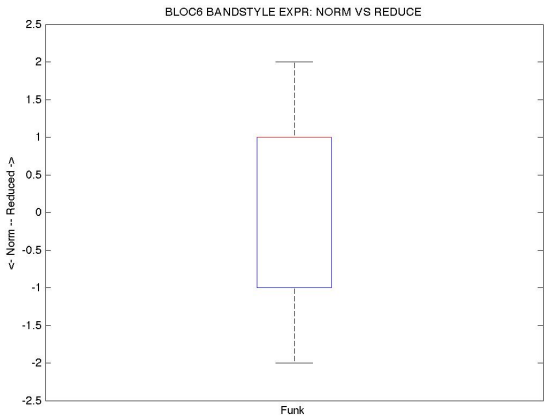


Figure 5.24: Bloc6 BandStyle Expr: Norm vs Reduced

The word “Funk” presents seems to be a non-realiable word, because in the synth process, as it is not a spanish word, the las phoneme has been erased. Eventhough, we can se in the results that depending on the model the results are not very coherent.

BandInfo/RadioInfo results

In that case we have only compared the two enhanced models in order to have and approach of which one could be better than the other. We have to remember that the tags included in the Freq models are also in the Reduced model, but the last one includes also additional tags (see in 4.2.1).

Here we can see the results for the tag Band.info and Radio.info with the word “Omar Rodriguez” and “Ernesto Soldevilla” for the CALM speaking style models:

Figure 5.25: Bloc7 BandInfo Calm: Freq vs Reduced

and here we can se the results for the EXPRESSIVE speaking style models:

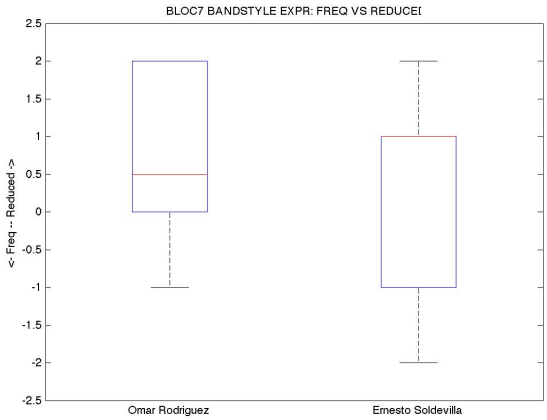


Figure 5.26: Bloc7 BandInfo Expr: Freq vs Reduced

In both cases the Reduced models (with the BandInfo tag added) have better ratings, so we can guess that this clustered tags of the Reduced dataset (4.2.1) enhance the created models.

Conclusions

Once we have all the results we can extract two main conclusions:

The first one is that the enhancement strategy seems to work, as in the most of the cases the enhanced (Freq and Reduced) models get the better results vs the Norm model. Even that, not always the addition of the tags enhance the specific tagged word. As seen in the

test, where sometimes the Norm models of a single word gets betters results, it seems that the inclusion of the tags enhance the whole average of the sentences and not the tagged words. So, even the results do not show a clear superiority of the enhanced models, we can see that the our approach is correct, but a biggest data collection will be necessary to corroborate the results.

On the other hand, even the results shows that the Reduced model performs little better than the Freq model, it would be great the perform other test to verify this small results.

6 Future work and conclusions

6.1 Future Work

This section summarizes the research work that could be done to improve the current work. Also, from an application perspective, some ideas are given. From the research point of view, the following aspects could be taken into account:

- Record more training data in order to obtain better models. The current models uses a total of 259 (calm)/ 244 (expr) sentences. This is probably not enough to generate a reliable model. Recording other speakers or more from the same one would help to increase the amount of training data. This new training data should be include only spanish names in order to avoid the transcription problems that have influenced in the final result. Also, include more labels (context information) in this new data.
- Other ways of adaptation the HTS to Vocaloid parameters will be good to try, or a post-tunning process in order to get better sounding sentences.
- Develop the Natural Language system to create the radio shows dynamically.
- Develop a system able to collect info from the web related to the songs in the playlist. This information will be able to be included into the created radio shows if the system need it.
- Finally, the technologies and codes used in this project belong to different operating systems. It would be a helpful and faster to get the results the use of an integrated framework under the same operating system.

6.2 Conclusions

The main conclusion of this project is that the proof of concept of what an expressive speech synthesis for a Radio DJ could be has been set. Even that, still remains lots of work to do concerning data collection, labeling, models enhancement and processing. Different conclusions can be considered regarding the different parts of the project, such the corpus design, the methodology, the adaptation process, the results as well as the evaluation process.

Concerning the Radio DJ corpus it has been build looking to common sentences and structures according to the presented songs and real radio show examples. The important

information related to this songs has been labeled. This information has been included during the creation of the radio shows without taking into account the frequency of appearance. Only thinking in the coherence of the shows. This gives more naturalness to the shows and it is useful in order to have templates where the tagged words can be interchangeable. Otherwise, in terms of modeling perhaps would have been better to create radio shows with more utterances of tagged words (context information). Another important point in the creation process of the Copus is the inclusion of non-spanish words into it. It has been a problem in some cases. Some of the phonemes of the english words don't exist in the spanish phonetics. For further works it has to be taken into account to avoid future problems.

Regarding the methodology followed to create de Radio Dj expressive models, more data would be desirable to train the models properly. This data could come from the same radiospeaker (semiprofessional), but could be interesting to record data from other professional radiospeakers too. Even that, it is difficult to give to the speaker concise instructions concerning how to express each emotion and also is difficult to the speaker to maintain the same expressivity during 6 radioshows. The labeling process has been done by clustering the same kind of context information into the same tag. Depending on the strategy some tags have been excluded and others clustered in a global-common info tag (such Bandinfo, Radioinfo). As seen in the results the strategies enhance the models. However, could be interesting to test other types of clustering or at least have more utterances per tag.

The adaptation process from HTS to Vocaloid, as it is not a direct process, has some issues that add to the final synthesized audios some artifacts. For example some tremor in the voice and abrupt falls in the pitch at the end of the sentences. Some feedback given by the evaluators denotes this artifacts that make difficult the focus on the expressiveness without taken into account them. It would have been good to spend more time in the processing process of this part. However it produces better quality audios than the ones generated with the HTS spectral content models.

Concerning the evaluation process. We had perform a evaluation test to people without familiarity with speech synthesis. This could give us a non-biased point of view of the results. Some other feedback received from the evaluators is that sometimes the difference are to tiny and some times to relative. As seen in chapter 5, pairs of audio of different models have been presented having as a rating process a five-point scale. We tried to perform a test with all the possible combinations that can answer the hypothesis of this master thesis. As the test cannot be to long because the people gets tired, we could not do all the pairs comparisons that we would like to do. Above all, the ones that compares the both enhanced models. About the results obtained from the test, we can convey that they denote that the approach is well directed. Although still remains lots of work to do.

Bibliography

[AL05] Alonso, M., Expressive Performance Model for a Singing Voice Synthesizer, Master thesis, MTG-UPF, 2005.

[AN12] Sebastian Andersson, Junichi Yamagishi, Robert A.J. Clark, Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis, Speech Commication, Volume 54, Issue 2, February 2012, Pages 175-188, ISSN 0167-6393, 10.1016/j.specom.2011.08.001.

[AY08] Aylett, M., Yamagishi, J., 2008. Combining statistical parametric speech synthesis and unit-selection for automatic voice cloning. In: Proc. LangTech.

[BA09] Barra, R., Yamagishi, J., King, S., Montero, J. M., & Macías, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Communication, 52(5), 394-404.

[BE06] Bennett, C., Black, A., 2006. Blizzard Challenge 2006. In: Proc. Blizzard Challenge Workshop.

[BL97] Black, A., Taylor, P., 1997. Automatically clustering similar units for unit selection in speech synthesis. In: Proc. Eurospeech. pp. 601–604.

[BL02] Black, A., 2002. Perfect synthesis for all of the people all of the time. In: Proc. IEEE Speech Synthesis Workshop.

[BL03] Black, A.W., 2003. Unit selection and emotional speech. In: Proc. EUROSPEECH 2003, pp. 1649–1652.

[BL05] Black, A.W., Tokuda, K., 2005. The Blizzard Challenge – 2005: evaluating corpus-based speech synthesis on common datasets. In: Proc. EUROSPEECH 2005, pp.77–80.

[BON07] Bonada, J., Serra, X., "Synthesis of the Singing Voice by Performance Sampling and Spectral Models", IEEE Signal Processing Magazine, vol. 24, no. 1, pp. 69-79, March 2007, ISSN 1053-5888.

[DA1872] Darwin, C. (1872). The expression of emotion in man and animals. London: John Murray.

[DJ1791] Degen, J., Mechanism of the human speech with description of its speaking machine. [DO95] Donovan, R., Woodland, P., 1995. Improvements in an HMM-based speech synthesiser. In: Proc. Eurospeech. pp. 573–576.

[FL72] Flanagan, J., Speech Analysis, Synthesis, and Perception. Springer-Verlag,Berlin-Heidelberg-New York, (1972).

[FL73] Flanagan J., Rabiner L. (Editors) (1973). Speech Synthesis. Dowden, Hutchinson & Ross, Inc., Pennsylvania.

[FR07] Fraser, M., King, S., 2007. The Blizzard Challenge 2007. In: Proc. BLZ3-2007 (in Proc. SSW6).

[FRA05] Francesc Alías, Ignasi Iriondo, Lluís Formiga, Xavier Gonzalvo, Carlos Monzo, Xavier Sevillano; "High quality Spanish restricted-domain TTS oriented to a weather forecast application" , The 9th European Conference on Speech Communication and Technology (InterSpeech) ISSN 1018-4074, pp. 2573-2576, Septiembre, 2005, Lisboa (Portugal).

[FRE12] Freixes, M., Rapaloid: Adapting an Existing Framework for Emotional Speech Prosody Generation, Master thesis, MTG-UPF, 2012.

[HDSS06] History and Development of Speech Synthesis, Helsinki University of Technology, Retrieved on November 4, 2006.

[HI04] Hirai, T., Tenpaku, S., 2004. Using 5 ms segments in concatenative speech synthesis. In: Proc. ISCA SSW5.

[HO04] Gregor, O., Hofer,, 2004, Emotional Speech Synthesis, Master of Science School of Informatics University of Edinburgh.

[HU96] Hunt, A., Black, A., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP. pp. 373–376.

[KA04] Kawai, H., Toda, T., Ni, J., Tsuzaki, M., Tokuda, K., 2004. XIMERA: A new TTS from ATR based on corpus-based technologies. In: Proc. ISCA SSW5. [KAR08] Karaiskos, V., King, S., Clark, R.A.J., Mayo, C., 2008. The Blizzard Challenge 2008. In: Proc. Blizzard Challenge Workshop 2008, Brisbane, Australia.

[KI09] King, S., Karaiskos, V., 2009. The Blizzard Challenge 2009. In: Proc. The Blizzard Challenge Workshop, Edinburgh, U.K

[KL87] Klatt, D.,1987, Review of Text-to-Speech Conversion for English. Journal of the Acoustical Society of America, JASA vol. 82 (3), pp.737-793.

[KR08] Krstulovic, S., Latorre, J., Buchholz, S., 2008. Comparing QMT1 and HMMs for the synthesis of American English prosody. In: Proc. Speech Prosody. pp. 67–70.

[KTH06] Advances in Cognitive Psychology 2006, volume 2, no 2-3, 131-143, "The KTH synthesis of singing", Johan Sundberg Department of Speech, Music and Hearing, KTH, Stockholm.

[LB04] A. Loscos and J. Bonada. Emulating rough and growl voice in spectral domain. In International Conference on Digital Audio Effects (DAFx'04), Naples, Italy. Citeseer, 2004.

[LI06] Ling, Z.-H., Wang, R.-H., 2006. HMM-based unit selection using frame sized speech segments. In: Proc. Interspeech. pp. 2034–2037.

[OK06] Okubo, T., Mochizuki, R., Kobayashi, T., 2006. Hybrid voice conversion of unit selection and generation using prosody dependent HMM. IEICE Trans. Inf. Syst. E89-D (11), 2775–2782.

[MO90] Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9,453–467.

[NO07] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," IEICE Trans. Inf. & Syst., vol. E90-D, no. 9, pp. 1406–1413, Sept. 2007.

[NO10] T. Nose and T. Kobayashi, "A technique for estimating intensity of emotional expressions and speaking styles in speech based on multiple- regression HSMM," IEICE Trans. Inf. & Syst., vol. 93, no. 1, pp. 116– 124, 2010.

[NO11] Takashi Nose, Takao Kobayashi, Recent Development of HMM-Based Expressive Speech Synthesis and Its Applications, Proc. 2011 (APSIPA ASC 2011), Xi'an, China, 2011.

[PI06] Pitrelli, J., Bakis, R., Eide, E., Fernandez, R., Hamza, W., Picheny, M., 2006. The IBM expressive text-to-speech synthesis system for Amer- ican English. IEEE Trans. Speech Audio Process. 14 (4), 1099–1108.

[PL98] Plumpe, M., Acero, A., Hon, H.-W., Huang, X.-D., 1998. HMM-based smooth-ing for concatenative speech synthesis. In: Proc. ICSLP. pp. 2751–2754.

[PRP05] J. Posner, J.A. Russell, and B.S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development,

and psy- chopathology. Development and Psychopathology, 17(03):715–734, 2005.

[RO05] Rouibia, S., Rosec, O., 2005. Unit selection for speech synthesis based on a new acoustic target cost. In: Proc. Interspeech. pp. 2565–2568.

[ROM10] Romportl, J., Zovato, E., Santos, R., Ircing, P., Relano Gil, J., Danieli, M., 2010. Application of expressive TTS synthesis in an advanced ECA system. In: Proc. SSW7, Kyoto, Japan, pp. 120–125.

[RUS80] J.A. Russell. A circumplex model of affect. Journal of personality and social psy- chology, 39(6):1161–1178, 1980.

[SAI06] K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, HMM-based singing voice synthesis system, Proc. Interspeech, pp.1141-1144, Sept. 2006.

[SCH95] K.R. Scherer. Expression of emotion in voice and music. Journal of Voice, 9(3):235– 248, 1995.

[SCHR93] Schröder M. (1993). A Brief History of Synthetic Speech. Speech Communication vol. 13, pp. 231-237.

[SCHR01] Schröder,M., Emotional Speech Synthesis: A Review, DFKI, Saarbrücken, Germany, Institute of Phonetics, University of the Saarland

[SP08] Springer HandBook of Speech Processesing, Benesty, Jacob; Sondhi, M. M.; Huang, Yiteng (Eds.)2008, XXXVI.

[TA06] Taylor, P., 2006. Unifying unit selection and hidden Markov model speech syn- thesis. In: Proc. Interspeech. pp. 1758–1761.

[TA09] Taylor, P., Text-to-Speech Synthesis, University of Cambridge, 2009.

[TAC05] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, “Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing,” IEICE Trans. Inf. & Syst., vol. E88- D, no. 11, pp. 2484–2491, Nov. 2005.

[TAC06] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, “A style adaptation technique for speech synthesis using HSMM and supraseg- mental features,” IEICE Trans. Inf. & Syst., vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[TH06] Theune, M., Meijs, K., Heylen, D., Ordelman, R., “Generating expressive speech for storytelling applications”, Audio, Speech, and Language Processing, IEEE Transactions on, 2006, 1137 - 1144

[TR08] Traum, D.R., Swartout, W., Gratch, J., Marsella, S., 2008. A virtual human dialogue model for non-team interaction. In: Dybkjaer, Laila, Minker, Wolfgang (Eds.), Recent Trends in Discourse and Dialogue. Springer, Antwerp, Belgium, pp. 45–67.

[UM10] Umbert, M., Emotional speech synthesis for a Radio DJ: corpus design and expression modeling, Master thesis, MTG-UPF, 2010.

[WG04] G. Widmer and W. Goebl. Computational models of expressive music performance: The state of the art. Journal of New Music Research, 33(3):203–216, 2004.

[WO00] Wouters, J., Macon, M., 2000. Unit fusion for concatenative speech synthesis. In: Proc. ICSLP. pp. 302–305.

[YA06] Yang, J.-H., Zhao, Z.-W., Jiang, Y., Hu, G.-P., Wu, X.-R., 2006. Multi-tier non- uniform unit selection for corpus-based speech synthesis. In: Proc. Blizzard Challenge Workshop.

[YAM03] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis", Proc. EUROSPEECH 2003, vol.3, pp.2461-2464, Geneva, Switzerland, Sept. 2003.

[YAM05] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol. E88-D, no. 3, pp. 503–509, 2005.

[YO99] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," in Proc. of Eurospeech, 1999, pp. 2347–2350.

[YO00] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system", J. Acoust. Soc. Jpn. (E), vol.21, pp.199-206, April 2000.

[ZE07] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2007d. A hidden semi-Markov model-based speech synthesis system. IEICE Trans. Inf. Syst. E90-D (5), 825–834.

[ZE09] Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Comm. 51 (11), 1039–1064.

APPENDIX A

BAND

’bandname’: 92,

’bandstyle’: 25,

’bandcountry’: 7, -> bandinfo

’bandmembernameinstrument’: 6, -> bandinfo

’bandmembername’: 6, -> bandinfo

’bandleadername’: 4, -> bandinfo

’bandgentilic’: 4, -> bandinfo

’bandtourtime’: 3, -> bandinfo

’bandleaderinstrument’: 2, ’ -> bandinfo

bandleader’: 1, -> bandinfo

’bandsytle’: 5, -> bandinfo

’bandwebsite’: 1, -> bandinfo

RADIO

’radiospeakername’: 7, -> radioinfo

’speakername’: 3, -> radioinfo

’radioname’: 6, -> radioinfo

SONG

’songname’: 87, ’

’songtype’: 5, -> song.info

’songcomposername’: 4, -> song.info

’songyear’: 1, -> song.info

’composername’: 1, -> song.info

’songadjectives’: 1, -> song.info

ALBUM

’albumname’: 24,

’albumdiscographynumber’: 6, -> albuminfo

’albumepoc’: 3, -> albuminfo

’albumrecordingplace’: 1, -> albuminfo

’albumyear’: 1-> albuminfo

OTHER

’timezone’: 1, -> other

’linkdownload’: 3 , -> other

’time’: 5, -> other

’concertcitydate’: 7, -> other

APPENDIX B

SHOW 1

[SONG 1] Saludos, muy buenas tardes a todas, estamos en Música para pensar. Al empezar el programa hemos escuchado Chris Potter y su tema Train.

Seguimos en America para dar paso a David Fiuczynski con su tema Cherry Red. [SONG 2].

Cherry Red de Fuze. Damos paso ahora al Jazz de John Coltrane con el tema Giant Steps [SONG 3].

Seguimos con más Jazz [SONG 4]. Ahí estaba Kenny Garrett con la canción Alabama del disco Pursuance: The Music Of John Coltrane. Si te ha gustado toma nota! El día 17 de febrero de 2013 estará en la sala Luz De Gas de Barcelona.

[SONG 5] Lo que acabamos de escuchar lleva como título The Crunge. Esta canción es una grabación del saxofonista Joshua Redman del 2007 incluida en su disco Momentum.

Y seguimos con la magia sonora que desprenden canciones como esta, TUTU de Miles Davis [SONG 6]. Genial composición la de TUTU.

Cambiamos de aires y vamos con el Rock Progresivo de The Mars volta, y la canción Goliath de su álbum The Bedlam In Goliath [SONG 7].

Volvemos al Jazz, la siguiente canción El replicante es de los Madrileños 24 Frames que después de 9 meses de gira, dan su último concierto en Barcelona, 24 frames de El Replicante. [SONG 8].

Seguimos! con Killer Cat de Trentemoeller! [SONG 9].

Después de éste contundente Killer Cat, nos vamos con Ornette Coleman y su disco The Shape of Jazz to come, el cual contiene este corte titulado Lonely Woman. [SONG 10]

Vamos con un poco de Groove. [SONG 11] Locked in the Basement es el título de este disco de Heernt, del cual acabamos de escuchar Pi.

Seguimos con más Groove [SONG 12] A Contra Groove, uno de los clásicos de APAC. Si te han gustado puedes descargarte el álbum de APAC en www.apacband.bandcamp.com.

Esto va llegando a su fin! Os dejamos con un tema de FLY llamado Child's Play para acabar. [SONG 13]

Espero que te haya gustado la sesión de hoy. Soy Héctor y esto ha sido Música para pensar.

SHOW 2

Bienvenido a tu radio!, hoy empezamos con My Generation de The Who. [SONG 1]. Ahí estaba el tema compuesto por Pete Townshend de The Who, My Generation. Este tema es todo un icono de los sesenta.

Seguimos con más música de los sesenta. La canción que vamos ha escuchar ahora es de Yes, que da nombre al álbum Time and a Word. [SONG 2]

Seguimos con más música de la mano de Metallica, Master of Puppets!. [SONG 3].

Más música de los incombustibles Metallica, Welcome Home Sanitarium. [SONG 4].

Cambiamos de estilo y nos vamos con un poco de pop. Aquí están Oasis capitaneados por Los Hermanos Gallager, y su tema Wonderball. [SONG 5]

Es el momento de escuchar a Muse, con su nuevo trabajo que han titulado The Second Law [SONG 6]. Supremacy, eran Muse.

Hemos llegado ya al ecuador del programa. Saltamos de nuevo hacia el Rock Progresivo y damos paso a otra gran formación: Pink Floyd. De la cuál escucharemos su Money, Pink Floyd!. [SONG 7].

Ahí estaba el tema compuesto por Roger Waters, Money. Son las diez y media y continuamos con los ingleses, Pink Floyd. [SONG 8]. Precioso este "Whish you where here".

Seguimos ahora con Mark Turner y su disco YAM YAM del cual vamos a escuchar el tema Moment's Notice. [SONG 9].

Ahí la música de Mark Turner. Estamos llegando al final del show, pero aun nos queda tiempo para algunas canciones más como Groovin' High de Dizzy Gillespie. [SONG 10]

La siguiente banda es de Nueva York y se llaman Little Women. Podéis saber más sobre little women en weirdestbandintheworld.com. Son Little Women y su "Throat One". [SONG 11].

Esto ha sido todo por hoy! Muchas gracias por vuestra atención, espero que hayas disfrutado y nos vemos en la próxima!

SHOW 3

Agárrate que comenzamos! Soy el Ovejo y esta es tu radio!. [SONG 1]

Acabamos de escuchar Funk de Hysteriofunk. Si eres un fan del funk estás de suerte porqué Hysteriofunk está de gira!. Toma nota de esta fecha!: 29 de Marzo Hysteriofunk en directo en Barcelona.

El siguiente tema que vamos a escuchar forma parte del disco Maggot Brain de los enormes Funkadelic, estamos hablando ni más ni menos que de Super Stupid! [SONG 2].

Hoy la cosa va de Funk y por tanto no podía faltar James Brown! De su álbum Sex Machine, aquí tenemos Please Please Please! [SONG 3].

[SONG 4] Acaba de sonar Soul Testify una canción de los Zaragozanos The Faith Keepers. Vamos a poner otra canción de The Faith Keepers ya que este año ha sacado un nuevo disco titulado Leap Of Faith, The World Can Be So Mean de The faith Keepers!. [SONG 5]

Geniales estos jovencísimos The Faith Keepers! Seguimos con más funk! pues no podían faltar los míticos Tower of Power. De su álbum homónimo publicado en mil novecientos setenta y tres, Soul Vaccination, Tower of Power. [SONG 6]

[SONG 7] Hemos escuchado Chameleon de Herbie Hancock. Chameleon es una de las cuatro canciones que componen el álbum Head Hunters.

Nos movemos ahora hacia el soul de la mano de Otis Redding [SONG 8] Just One More Day tituló en su momento Otis Redding esta canción.

Y siguiendo con la tónica de grandes canciones del Funk, aquí tenemos a Soulive con su temazo Tuesday Night Squad. [SONG 9].

Seguimos con más música de la mano de Stevie Wonder y su tema Superstition. [SONG 10]. Genial este Superstition.

Os dejamos con una versión del tema Sing Sing de Seamus blake.

Espero que te haya gustado la sesión de hoy. Soy el Ovejo y esto ha sido La Jam Session. [SONG 11]

SHOW 4

[SONG 1] Bienvenidos a una nueva sesión de música con Cifu, esto es A todo Jazz. Hemos empezado con un tema llamado Autumm Leaves, interpretado por el quinteto de Miles Davis del álbum grabado en directo en el Village Vanguard.

Continuamos con más música del quinteto de Miles Davis. El siguiente tema forma parte del disco Kind of Blue de Miles Davis y se llama Blue in Green.

El tema fue compuesto por pianista de la banda Bill Evans. Ahí va Blue in Green. [SONG 2] Preciosa balada Blue in Green.

En el Quinteto de Miles Davis había músicos brillantes como John Coltrane o Paul Chambers los cuales se juntarían posteriormente en otras formaciones como la que grabó el álbum Soultrane de John Coltrane, del cual vamos a escuchar la versión Good Bait [SONG 3].

Damos paso ahora a otra gran formación de la época: El quinteto de Cannonball Adderley, y el blues Spontaneous Combustion grabado en el disco de presentación de Cannonball [SONG 4]

[SONG 5] Ahí estaba el tema compuesto por Billy Strayhorn, Isfahan, interpretado por el saxofonista Joe Henderson.

Continuamos con más música de Joe Henderson de su álbum In'n'Out, publicado en el sesenta y cuatro, Punjab se llama el tema, ahí está. [SONG 6]. Son las tres menos cuarto y continuamos con uno de los muchos discos que grabó el trío compuesto por Brad Meldhau, Jorge Rossy y Larry Grenardier bajo el nombre The Art of Trio. El tema se titula All the things you are [SONG 7]. All the things you are, uno de los clásicos del jazz.

Volvemos al Quinteto de Miles Davis, con el tema All Blues. [SONG 8] Precioso este All Blues con Bill al piano soberbio!

[SONG 9] La composición que acabamos de escuchar lleva como título Mr PC. Éste tema fue compuesto por John Coltrane en dedicatoria a su contrabajista Paul Chambers.

Se no echa el tiempo encima, vamos a acabar con un tema de Llibert Fortuny llamado Double Step. Muchas gracias por vuestra atención, disfruta y nos vemos en la próxima!. [SONG 10]

SHOW 5

Bienvenidos a una nueva sesión de música con Benja, esto es La Oveja Negra Rock. Hoy vamos a comenzar nuestro programa con Roots Bloody Roots de Sepultura. [SONG 1]. Ha sonado Roots Bloody Roots de los Sepultura.

[I F T T]Continuamos con el Metal Llegados desde Finlandia, aquí tenemos a Throne of Chaos. [SONG 2]. Contundente como el que más éste tema llamado Pervertigo.

Nos vamos al país vecino con los Suecos Freak Kitchen y su Hard Rock liderado por la magnífica guitarra de Mattias. Freak Kitchen, Speak when Spoken to! [SONG 3]

Y de su primer álbum Appetizer, aquí tenemos Blind, también de los Freak Kitchen! [SONG 4]

[SONG 5] Hemos cambiado de país y de subgénero pero seguimos con el Metal! El tema era Orion de Metallica.

Si eres un fan del $band.style estás de suerte porqué $band.name están de nuevo de gira! Podéis saber más sobre el tour de Metallica en su página web oficial.

Seguimos en La Oveja Negra Rock, son las 12:00h y vamos a por otra instrumental de Metallica, To Live is To Die [SONG 6]. Y con esta increíble To Live is To Die hemos llegado ya al ecuador del programa.

Vamos a escuchar ahora bandas del panorama actual del Metal. Del álbum Leviathan de Mastodon tenemos Blood and Thunder! [SONG 7].

Y siguiendo con la tónica de grandes grupos del panorama actual vamos con Shot By Bertha y una par de temas de su primer disco Hangover. [SONG 8] [SONG 9]

Massive Collapse y Rule Number one son los nombres de las dos canciones que hemos escuchado, canciones de blandenses Shot By Bertha.

Y de un grupo novel pasamos a una de sus grandes influencias, como son Machine Head. En su disco The Blackening podemos encontrar temazos como este Clenching the Fists of Dissent [SONG 10].

Es momento de dar unos minutos a unos clásicos del Metal, como lo son Judas Priest y su mítico Painkiller [SONG 11]

Acabamos el programa de hoy con otro tema instrumental de Metallica, The Call of the Ktulu [SONG 12]

Un saludo de Benja al micro, nos vemos en otra sesión de La Oveja Negra Rock, hasta la próxima.

SHOW 6

Bienvenidos a una nueva sesión de música con Ernesto Soldevilla, esto es Radio A todas Horas. Hoy vamos a comenzar nuestro programa , el cual será muy variado, con algo de Pop, y será con She will be loved de Maroon5. [SONG 1].

Seguimos con más Pop, La siguiente banda que vamos a escuchar es de Alemania y se llaman Xavier Naidoo. [SONG 2].

Nos pasamos al lado más rockero del pop para escuchar a The White Stripes y la canción, que también titula su último trabajo, Icky Thump [SONG 3].

Después de éste genial Icky Thump vamos a escuchar otro tema de uno de los muchos proyectos que tiene el guitarrista Jack White.

El tema se titula Treat Me Like Your Mother y forma parte del proyecto The Dead Weather [SONG 4].

Vamos ahora por uno de los éxitos del 2012, el Gangnam Style! [SONG 5].

Y de éxito a éxito! Ahí un tema que triunfó en los 80s, estamos hablando de Thriller del ya difunto Michael Jackson [SONG 6].

Hacemos un cambio radical y saltamos ahora hacia el Tecno!. [SONG 6]

Hemos escuchado el último éxitazo de David Guetta, llamado I Gotta Feeling.

Ya se nos ha ido medio programa pero aún tenemos tiempo para muchas más canciones [SONG 8].

La canción que acabamos de escuchar lleva como título Ich will y forma parte del tercer álbum de los también Alemanes Rammstein titulado Mutter.

Y de Alemania nos vamos cerca, a Francia donde los Ulan Bator crean su Post-Rock, el tema que vamos a poder escuchar hoy aquí se titula Fuiete [SONG 9].

Seguimos con el Post-Rock, de la mano de At the Drive-In [SONG 10].

Lo que acabamos de escuchar lleva como título Napoleon Solo y forma parte del disco In/Casino/Out de los At the Drive-In.

El guitarrista de AT the Drive In, Omar Rodriguez Lopez es un creador inagotable de ideas, y aquí os vamos a poner un par más de ellas antes de encarar la última parte del programa, ahí van! [SONG 11] [SONG 12]

The Power of Myth y Electrodorphines son los nombres de las dos canciones que hemos escuchado, canciones de uno de los muchos grupos de Omar Rodriguez.

Encaramos la recta final del programa con un cambio de estilo, vamos a por el Funk, Maceo Parker, Pass the Peas! [SONG 13] Acabamos con un tema de los jovencísimos Jizzy Free, Lets' Go [SONG 14] Un saludo de Ernesto Soldevilla al micro, hasta la próxima.

APPENDIX C

Context-dependent label format for HMM-based speech synthesis:

p1 the phoneme identity before the previous phoneme
p2 the previous phoneme identity
p3 the current phoneme identity
p4 the next phoneme identity
p5 the phoneme after the next phoneme identity
p6 position of the current phoneme identity in the current syllable (forward)
p7 position of the current phoneme identity in the current syllable (backward)
a1 whether the previous syllable stressed or not (0: not stressed, 1: stressed)
a2 whether the previous syllable accented or not (0: not accented, 1: accented)
a3 the number of phonemes in the previous syllable
b1 whether the current syllable stressed or not (0: not stressed, 1: stressed)
b2 whether the current syllable accented or not (0: not accented, 1: accented)
b3 the number of phonemes in the current syllable
b4 position of the current syllable in the current word (forward)
b5 position of the current syllable in the current word (backward)
b6 position of the current syllable in the current phrase (forward)
b7 position of the current syllable in the current phrase (backward)
b8 the number of stressed syllables before the current syllable in the current phrase
b9 the number of stressed syllables after the current syllable in the current phrase
b10 the number of accented syllables before the current syllable in the current phrase
b11 the number of accented syllables after the current syllable in the current phrase
b12 the number of syllables from the previous stressed syllable to the current syllable
b13 the number of syllables from the current syllable to the next stressed syllable
b14 the number of syllables from the previous accented syllable to the current syllable
b15 the number of syllables from the current syllable to the next accented syllable
b16 name of the vowel of the current syllable
c1 whether the next syllable stressed or not (0: not stressed, 1: stressed)
c2 whether the next syllable accented or not (0: not accented, 1: accented)
c3 the number of phonemes in the next syllable
d1 gpos (guess part-of-speech) of the previous word
d2 the number of syllables in the previous word
e1 gpos (guess part-of-speech) of the current word
e2 the number of syllables in the current word
e3 position of the current word in the current phrase (forward)
e4 position of the current word in the current phrase (backward)
e5 the number of content words before the current word in the current phrase
e6 the number of content words after the current word in the current phrase
e7 the number of words from the previous content word to the current word

e8 the number of words from the current word to the next content word
f1 gpos (guess part-of-speech) of the next word
f2 the number of syllables in the next word
g1 the number of syllables in the previous phrase
g2 the number of words in the previous phrase
h1 the number of syllables in the current phrase
h2 the number of words in the current phrase
h3 position of the current phrase in utterence (forward)
h4 position of the current phrase in utterence (backward)
h5 TOBI endtone of the current phrase
i1 the number of syllables in the next phrase
i2 the number of words in the next phrase
j1 the number of syllables in this utterence
j2 the number of words in this utternce
j3 the number of phrases in this utterence

APPENDIX D

To see the Perceptual test, please look at the attached documents.