University of Sheffield

Automatic Speech Recognition in Music

Gerardo Roa Dabike

Supervisor: Dr Jon Barker

A report submitted in fulfilment of the requirements
for the degree of MSc in Advanced Computer Science

in the

Department of Computer Science

September 6, 2016

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: Gerardo Roa Dabike

Signature:

Date: September 6, 2016

Acknowledgement

I have taken efforts in this research project. However, it would not have been possible without the kind support and guidance of my supervisor Dr. Jon Barker.

I would like to express my gratitude towards my wife Monserrat and children Catalina and Tomas for their kind encouragement and support which help me in completion of this project.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

Abstract

Automatic Speech Recognition in music is a barely analysed problem which can be beneficial in creative and retail business applications. This project is aimed to experiment in a musical corpus with synthetic augmented training data and with DNN methodologies in order to determine if these approaches can improve the performance of recognizer. Previous researches used speaker to singer adaptation rather than training in a singing database. First, creating a novel corpus ACOMUS1 based in acoustic cover music and then several audio augmentations experiments will be conducted in order to try to reach a high performance and obtain information that would lead to further researches. The experiment results obtained a poor performance reaching a 86% of WER using DNN sMBR nevertheless, ACOMUS1 corpus showed to have a high growth potential that would allow more researches on it.

Acronyms

AM	Acoustic Model
ASR	Automatic Speech Recognition
DFT	Discrete Fourier Transform
DNN	Deep Neural Networks
FFT	Fast Fourier Transform
fMLLR	Feature Space Maximum Likelihood Linear Regression
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LM	Language Model
MFCC	Mel Frequency Cepstral Coefficients
MIR	Music Information Retrieval
MLLT	Maximum Likelihood Linear Transform
MLP	Multi-Layer Perceptrons
OOV	Out Of Vocabulary
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
SAT	Seaker Adaptive Training
sMBR	state-level Minimum Bayes Risk
SR	Speech Recognition
SRE	Speech Recognition Engine
WER	Word Error Rate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, Automatic Speech Recognition is a very complex and important topic which is present in several applications in the daily lives of people. In addition, singing is an artistic expression that accompanying people in their activities every day. Therefore, use ARS system to automatically transcript lyrics would help in several application such as, a plagiarism detection system, a music collection lyrics transcription, a karaoke lyrics generator and English singing pronunciation tester. For these purpose, a DNN approach in addition to augmented training data techniques will be study in order to improve the performance minimizing the WER.

1.1 Project Context

For mankind, music always had been an important mechanism to transmit knowledge, history and emotions. Throughout time, the means of transmitting have evolve from troubadours and live presentations to analogue and digital portable distribution technologies. Nowadays, thanks to the digital technologies and especially the **Internet**, thousand of music recordings are shared everyday, with multiple platforms storing and allowing redistribution of audio and video recordings. Moreover, on Internet are available several platforms that allow to share that audio and video recordings. Taking advantage of these technologies, artists shared their personal personal **acoustic covers** by popular artist. All of these options are significantly increasing the availability of music collections.

Music, can be described separating it component and transcribe it in the proper notation such us music scale for the musicality and lyrics for the singer. Knowing the lyrics from a music collection it is desirable for several applications like covers and plagiarism detectors, MIR, Karaoke games, non-native English singer pronunciation checker, among others. Nevertheless, ASR in music oriented to transcribe a song are rarely studied and most of the researches are focused in **singer recognizer** [30, 51, 50], **audio-lyrics alignment** [19] or MIR applications using different approaches such as MIDI and speech recognition systems [14].

This project will be oriented in evaluate HMM-GMM and HMM-DNN approaches for lyrics transcription on a music corpus and will attempt to determinate if augmenting the training database with synthetic modifications are appropriate techniques for improve the results. In addition, will be focus on English acoustic cover musics with one background instrument, any other kind of music is beyond the scope of this research. Moreover, a vocal separation method will not be evaluate in this research.

1.2 Aim and Objectives

This project will focus on evaluate HMM-GMM and HMM-DNN methodologies for ASR on music database. The aim of the project is to evaluate if to use synthetic augmentation training data in addition with HMM-DNN approach can significantly increase the performance In order to achieve this goal a corpus based in covers of popular music will be designed and constructed. Using this corpus, this dissertation will examine firstly, if a HMM-DNN approach can significantly increase the performance compared with the HMM-GMM results. Secondly, it will be study if to using synthetic augmentation of the training data can also be used in order to improve the results.

1.3 Overview of the report

The overall structure of the study takes the form of seven chapters, including this introductory chapter.

Chapter 2: Deals with the current literature of ASR system in music. Starting with the singing characteristics and it similarities and differences with normal speaking. Finishing with an review of the procedure and technologies of a normal SRE system.

Chapter 3: Will explain the motivation for creating the ACOMUS corpus in its first version. Also, the design process and construction procedure will be detailed.

Chapter 4: Will describe how the database corpus will be organize for training, development and testing purpose. Moreover, this Chapter will describe the training data augmentation techniques that will be used for the following experiments.

Chapter 5: The analysis of the definitions and construction of the Language Model will be explained in this Chapter. In addition, several consideration taken on the Acoustic Model will be also detailed.

Chapter 6: In this Chapter, HMM-GMM and DNN's approaches and experiments both in the original base and increased base will be discussed.

Chapter 7: The final conclusions of this research and further works will be presented.

Chapter 2

Literature Survey

Over the past decade most research in ASR in music has emphasized mainly in MIR systems and lyrics alignment rather than lyrics transcription systems. Exist several researches that analyse different aspects of music and also, describes several characteristics of the singing voice. Nevertheles, researches on music transcription using ASR systems emerge in 2010 with Messaros and Virtanen [28] research.

In this chapter the existing literature about singing properties and ASR systems in music will be covered. This survey will be travelling from the basics of any SR system to the currents researches in ASR in music. This travel include some fundamentals about the speech and more specific about the *singing voice*.

2.1 A Little Speech About Speech

It seems that one of the most important sounds that humans can produce are the *voices sounds*. These are sounds that involve a physical reaction which was explained by Sundberg [46] as any sound that start with an airstream from the lungs that pass through the vocal cords making its vibrate and is finally filtered by the pharynx, the mouth and the noise cavities. The importance of the voice sound lies in that these sounds are crucial for speaking and singing. In addition to the voice sound exist the *unvoiced sounds* which are sounds that do not start with an airstream from the lungs and just occurs in the vocal or nasal cavities. The main characteristic of any unvoiced sounds is that the vocal cords do not vibrate during it production.

Some specific arrange of these two kind of sounds are founded in speech and singing, generating an acoustic code for communication composed by words and sentences. The existence of this acoustic code suggests that each word should be pronounced equally independently the speaker. Nevertheless, the reality is that the way that the sound is produced depend on several factors like personal voice organs characteristics, the pronunciation and the accent of the speaker.

The voice organs have different length and dimensions between male and female and also, between people of the same gender. These characteristics have a direct influence on the *timbre* of the speaker and the immediate consequence of the that the same vowel may sound different between speakers. One gender characteristic given by the differences in the vocal organs is that male voice produce a lower pitch than female voice. These occurs because the dimension of men's vocal cords are longer and thicker than the females ones.

Additionally, the articulation of the vocal organ during phonation is defined as a function of the frequency of the transferred sound. In detail, the *fundamental frequency* F_0 is determined by the tension of the muscle of the vocal cords and this frequency contribute to the *pitch* of the voice. Moreover, exist several peaks in the waveform that are multiples of F_0 and it called *pitch harmonics*. Furthermore, the shape of the vocal track act as a very complex resonator that suppressed some harmonics but also, enhanced others. This enhanced harmonics are called **formants** [31]. Figure 2.1 shows the spectrum of the harmonics of the vowel a. It is possible to see that the first formant F_1 corresponds to the second harmonic h_2. The next harmonics picks are the second and third formants.

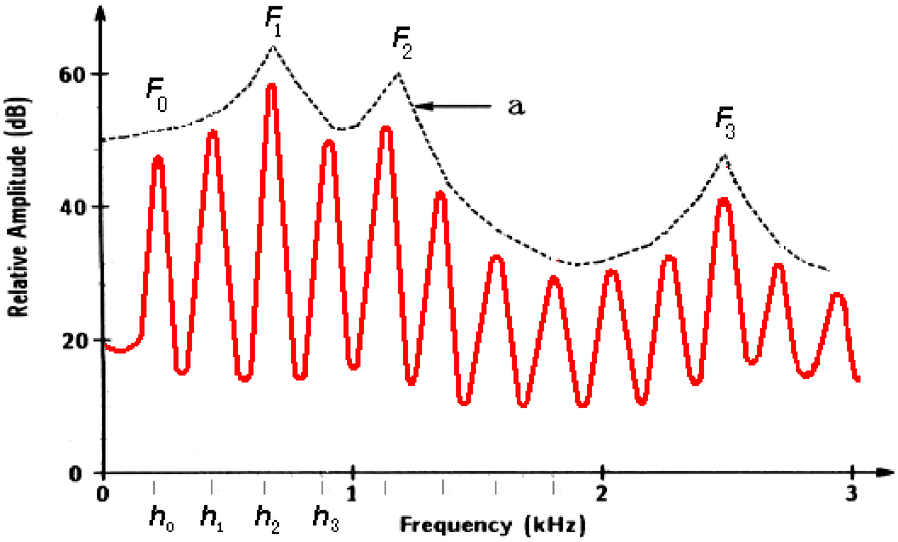
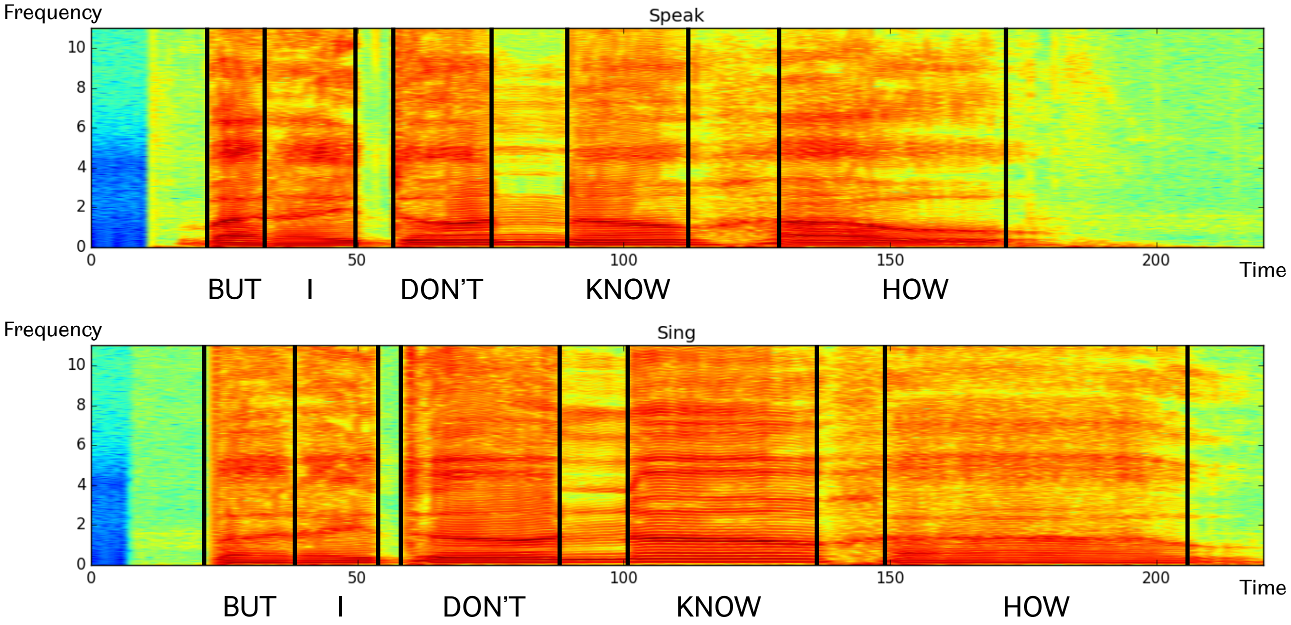


Figure 2.1: Spectrum of harmonics under the formants of vowel a.[1]

2.2 Singing Speech Characteristics

Some authors argue that music is one of the most complex audio signals because it carries much more information than other audio waveforms [10]. Exist a great collection of music in the world and it seems that the majority have singing voice as one of its components. The singing voice is a modification of the normal speech that depending the ability of the artist can be modify the way of each word is spoken. In addition, some authors [21] describe the singing voice as a very versatile instrument that can cover frequencies from two to five octaves. Therefore, it can be argue that the properties of the singing voice are more complex than the properties of the speech voice. Unlike speech, singing tends to focus on intonation and musicality rather than in the complete intelligibility of the words. Moreover, the fluctuation of the frequencies of the singing voice between one phoneme and another can be very soft and longer but more musically harmonic, this is common

[1] Source: From the Web site *The Cochlea* [?]



In both spectrograms the same sentence ”But I Don't Know How” from the song **Wonderwall** from the **Oasis** was spoken and singed. It is possible to see that in the singing sentence the vowels are longer and the transition between one phone to another are smoother than the spoken sentence, This different can be appreciate clearly in the words **KNOW** and **HOW**.

Figure 2.2: Spectrograms of a sung and spoken sentence.

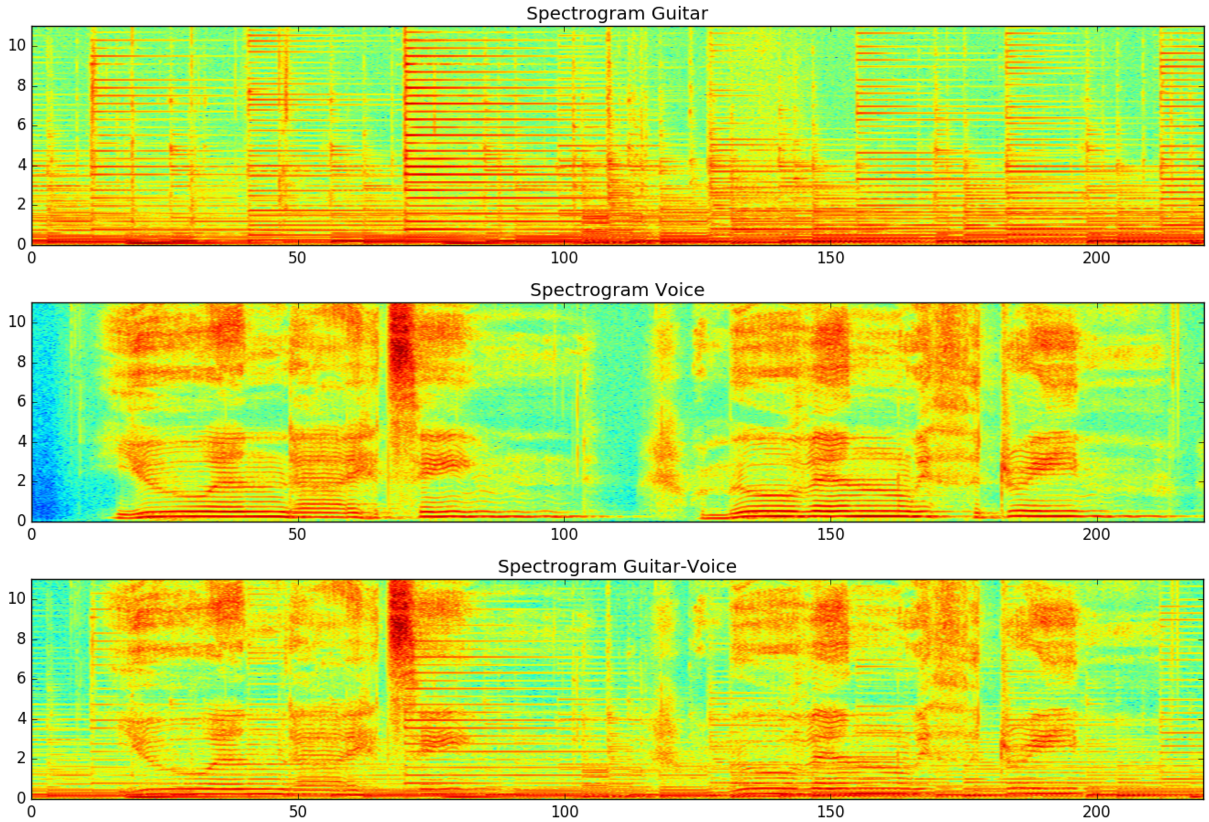
with the vowels which tend to be significantly longer than in normal speech. In contrast, the speech voice can change freely in pitch and loudness.

In order to illustrate the difference of existent in sing and speak, Figure 2.2 detail the spectrogram of the same speaking and singing sentence.

2.2.1 The Singing and the Instrument

In common ASR systems, the music in background is considered as a environmental noise sound [40]. Nonetheless, in music the speech is complemented by the accompaniment instrument and work together in order to keep harmonies of the song. Consequently, this close relation between the instruments and the speech add complexity and distortion in any speech recognizer. In addition, is expected that the sound of the instruments overlap with the speech frequencies. And also, this occur differently for each instrument because the frequency ranges of its. For instance, an acoustic guitar play notes between 82 Hz (E2) and 1.4 kHz (F6) and a piano play notes between 28 HZ (A0) and 4.1 kHz (C8) [56]. These instruments frequencies can overlap with the singer formants and most precisely, with the frequencies between the range of 2 Khz and 4 KHz. Figure 2.3 illustrate how a guitar affects the singing frequencies and the energy in each formant. A classical singer training tend to focus in voice enhancing, increasing the signal intensity between the third and fourth formants also called **singer's formant** [1]. These is the most important differentiation with normal speech in terms of the frequencies.

[2] Extraction from the track *LizNelson_Rainfall_MIX* from the sample database of [2]



*Notes:*In th guitar's graph each pitch and its harmonics are constants in each note. Moreover, in the case of the singing's graph it can be seen the long vowels that was previously explained and also, the smoothed transition between sounds. Furthermore, in the third graph the spectrogram of the waves of singing and guitar are combined. In this case, it is notorious that superposition of the guitar impacts the singing becoming the transition smoother than singing alone.

Figure 2.3: Comparative of the spectrograms of a guitar, a singing voice and its combination.[2]

In addition to singer's formant, authors classified others features that are very characteristic and important in singing voice and in music like the **Style**, the **Pitch**, the **Timbre** and the **Tempo** [10, 21, 46].

2.2.2 Styles

In general, the style in music represents a regional and social differences between singers. That is how, for example, the **rap** and the **punk** music identify a group of people with specific ideals but, the **folklore** music identify custom and specific geographic locations. It is possible to compare the style in music with the speakers accents in a spoken language because, is the accent that represents a regional or social characteristic [32]. In addition, the style is also connected with the register of the sound which can be defined as the kind of voice sound that the singer can produce in specific styles. Therefore, one can expect similar register for the same style of music independently the singer and also, one can expect that a singer can change the register depending the style that he is interpreting. Nonetheless, the changes in the register depend directly on the ability and the training of the artist. For instance, for a singer change the style from modal to falsetto can be done with a minor modification of the vocal track, whereas changing from modal to voix-mixte needs major modifications in pharynx, lips and mouth and therefore, mayor training [21]. In other words, potentially all the artist should be able to modify their interpretation between one style to another however, some of this modifications demands high training from the artist.

2.2.3 Pitch

In singing voice, the pitch range is usually higher than in speech voice [10] and its control is a very important skill in music. A singer can maintain the pitch or change it as his desires in order to interpret the song in a specific style or just for keep a note. In contrast, in normal speech the pitch varies all the time and there is no need to controlled.

In singing, the first and second formant have similar behaviour than normal speech. However, the third and forth formant have a big relevance in singing voice and the intensity between these two formants is called *singer formant* [1]. Moreover, the centre of the singing formant have harmonics peaks between 2 KHz and 4KHz depending of the singer's voice type. Specifically, this is 2.2 kHz in basses, 2.7 kHz in baritones, 2.8 kHz in in tenors and 3.2 kHz in altos [47]. The measure of the singer resonant quality is called **Singer Power Ratio** and is defined as the ratio between highest intensity peak in the range of 2 and 4 KHz and the highest intensity between 0 and 2 KHz [34]. Some authors claim that this measure is similar in sing and speech in training singers and therefore, can be used as a metric of the evolution of a training singers [26]. Nonetheless, all these studies are related with classical singers that follow the normal tones scales. Nonetheless, contemporary singers tend to sing in less restricted scales using more flat notes making more difficult to elaborate an objective definition.

Finally, despite the difference between the singing and speaking pitch, exist a very important constant that is the pitch ranges per male and female. As it is known, female have a higher pitch than males and this is the same in speaking as in singing. Taking advantage of this quality, in this research some experiments using variations of the pitch will be conducted. These experiment will be detailed in Chapter 4.

2.2.4 Timbre

The timbre is a property related with the musical note. Therefore, even when two instruments play the same pitch and loudness, the note will not sound equal because of the timbre of each instrument. In singing, the timbre in vowels depends on the formant frequencies which are controlled by the singer using the characteristics of their vocal track. This means that the singer's control of the vocal track produces a slightly difference between one singer and another. Moreover, there is also gender difference in the timbre for instance, the male voice spectrum has a weaker fundamental than the female voice spectrum [28]. It seems that this characteristic will not affect the accuracy of an ASR in music. Nevertheless, if the singer is talented the timbre might help to improve a singer dependant recognizer.

2.2.5 Tempo

Tempo is the characteristic of the music that is frequently associated to the expressive nature of the music. It is defined as *"The speed at which a piece of music is performed"*[22]. Some researches indicates that the tempo in music and its perception can not only, be determinate by the age of the listener, but also, by the style [41, 39]. Tempo can be represented numerically in *Beat Per Minute* and more typically in subjective terms such as *adagio* and *allegro* terms. But subjectively, listeners can identify the tempo of a piece of music as a *fast* or a *slow* song. This temporal property of the music will be also explored in this research and will be explained in Chapter 4.

2.3 ASR in music

The goal of a ASR research in music is the same that any ASR research, which is to build a computational systems that match a string of words with the audio signal [18]. The main different is that, as was mentioned before, a music wave-stream carry on more information than a read-speech or conversational-speech signals. A generalization of the ASR procedure used and it stages will be explained in this section.

2.3.1 Audio Signal Analysis

The first step in any ASR system is to transform the input waveform into a digital representation. Usually, this representation is in the form of a subsequence **features vectors** where each vector intent to identify the linguistic information in a small window of the signal [18]. There are many feature representations for an audio signal but, by far the most commonly front-end used in speech recognition is the **MFCC** based on the cepstrum [3] idea and this is the one that will be used in this project.

In this research the computation of the MFCC used will be the same that Kaldi follows [36]:

- Work out the number of frames in the file typically $25ms$ frames shifted by $10ms$ each time.
- For each frame:
  1. Extract the data, do optional dithering, pre-emphasis and dc offset removal, and multiply it by a windowing function.
  2. Work out the energy at this point.
  3. Do **FFT** and compute the power spectrum.
  4. Compute the energy in each mel bin.
  5. Compute the log of the energies and take the cosine transform, keeping as many coefficients as specified.
  6. Optionally do cepstral liftering; this is just a scaling of the coefficients, which ensures they have a reasonable range.

In addition to the static feature extraction, for some approaches are necessary and beneficial to use some transformations techniques in order to improve the results. One of these techniques are LDA transformation with MLLT estimation. On the one hand, LDA is a transformation matrix that reduce the feature dimensionality preserving most of the necessary information that can be used

[3] The cepstrum is defined as *"the inverse Fourier transform of the log magnitude spectrum of a signal"* [33].

for a class identification. This technique require that the classes are properly labelled [13, 43]. On the other hand, MLLT introduce a new form of a covariance matrix which allows sharing a few full covariance matrices over many distributions maximizing the likelihood of the training data [38, 43].

2.3.2 Gaussian Mixture Models

A **GMM** is composed by a combination of M Gaussians distributions where each Gaussian is weighted by it likelihood contribution 2.1.

$$p(x) = \sum_{k=1}^{M} w_i N(x, \mu_i, \sigma_i) \tag{2.1}$$

$$\text{where} \sum_{i=1}^{M} w_i = 1$$

Where each Gaussian is defined by Equation 2.2, where μ in the mean and σ^2 is the variance.

$$f(x\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \text{exp}\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \tag{2.2}$$

$$\mu = \sum_{i=1}^{N} p(X_i) X_i$$

$$\sigma^2 = \sum_{i=1}^{N} p(X_i)(X_i - E(X))^2$$

2.3.3 Hidden Markov Models

A **HMM** is composed by a sequence of states that have an GMM probability density function associated to each state. Moreover, a transition probability matrix is associated between the states. In the training stage, the transition matrix and GMM variables must be estimated in order to maximize the likelihood of each training observation vector.

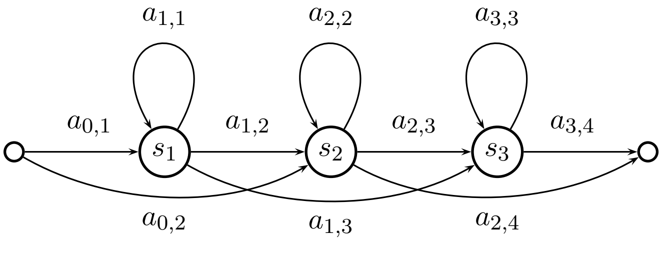


Figure 2.4: A Hidden Markov Model.[4]

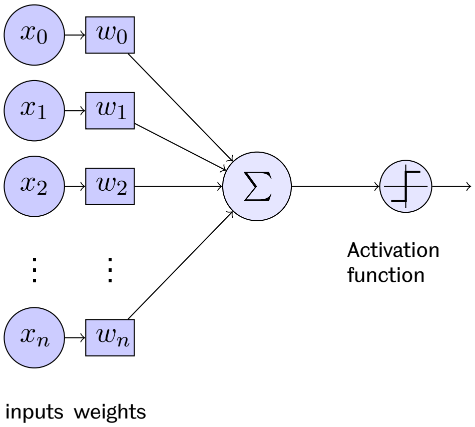
Figure 2.4 represents the possible flow of an HMM with three states plus start and end state. This diagram is typical in ASR where the probability of the start state is 100% and the flow direction

[4] Source: Diagram designed by [49]

is strictly left to right. There are two training algorithms for ML estimation in HMM models. The first one is the *Viterbi* algorithm which estimates the most likely path. This allows the Viterbi algorithm to avoid the problem of the hidden states. On the other hand, the *Baum-Welch* algorithm takes this hidden states into account taking the state posterior probability. The re-estimation algorithm can guarantee that the maximum local likelihood is obtained.

2.3.4 Deep Neural Networks

In the origins of the *Artificial Neural Networks* the motive was to mimic how the human brain works [42]. Research on this topic introduced the term *perceptrons* which suppose to be a machine rather than a program and it algorithm intent to decide if a feature vector belong to a class. An illustration of the perceptrons algorithm can be found in Figure 2.5. Here is shown how a neural unit from an input layer x_i is weighted by a w_i factor and then an *activation function* is applied.



$$output = \begin{cases} 1 & \text{if } \sum_{i=0}^{n} w_i x_i > 0 \\ \\ -1 & \text{otherwise} \end{cases} \tag{2.3}$$

Figure 2.5: Diagram of perceptrons algorithm

Multi-layer Perceptrons

Multi-layer Perceptrons approach were widely developed for unsupervised learning algorithms in the 1980's [44]. Figure 2.6 present the diagram that represents how a MLP connects the different existing layers in order to generate the output layer. Deep Neural Networks are MLP systems that instead of increase the number of layers they contain deep structures. Moreover, they work with multi frames and use undirected generative models called restricted Boltzmann machines [16].

In ASR system, the DNN training process consists in fitting the model with a chunk of up to 31 frames. The main idea is to send each frame to the model with the information about the next and previous frames. With this data the networks knows if the frequencies represented in that frame are going up, down or remain constant [15].

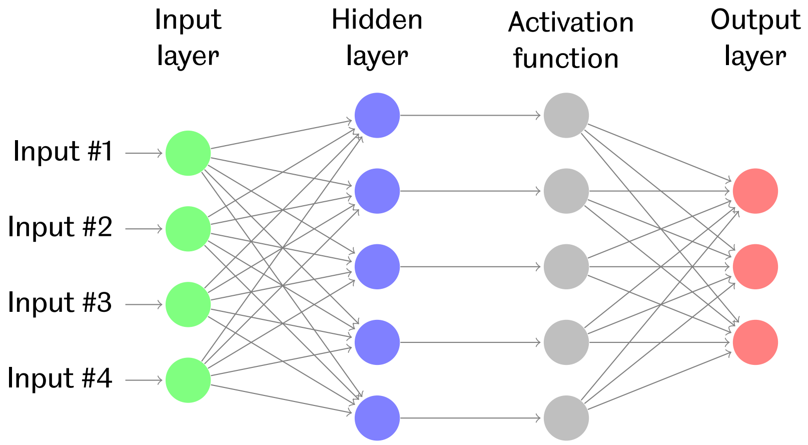


Figure 2.6: Multi-Layer Perceptron

2.3.5 Singer separation

For humans and machines, identify an specify source in a multiple sources signal it is very difficult and one of the possible solutions to this problem is to isolate the required source. Nevertheless, the process of separation has its own difficulties due to two main reasons defined by Weiss [53].

1. **Indistinctness** of the signal: It is the unreliability of identify discrete harmonics in Furier transform.
2. **Interference**: It is the difficulty of separate and identify the origin of frequencies when are too close between each other [10].

Existe several studied that investigate methods to identify the singers from a musical signal. Some authors investigate HMM model using factorial HMM methods obtaining 58% of accuracy in source-adapted models and 62% in speaker dependant models [10]. In contrast, other authors investigate a separation using the time-varying pitch enhancing and subtracting the background music. Speech enhancement is one of the most important topics in speech recognition in noisy environments. Several techniques that address this topic exist, like the *spectral subtraction* approach which estimates the spectrum of clean signal by subtracting of the estimated noise magnitude spectrum [3]. Another technique is the *signal subspace* approach that focuses on enhancing a speech signal degraded by an uncorrelated noise [12]. *Weiner filter* is the most popular technique that had been widely used [9]. This approach depends on the adaptation of the filter function from sample to sample based in the statistical parameters $mean$ and $variance$. The performance of each methodology depend directly on the quality of the signal [6].

2.3.6 Evaluation

In general, any ASR system is evaluated using the metric **WER** which represent the minimum distance between the hypothesized and correct string. Exist three different classifications of what are considered as wrong words which are: the *deletions* or words missed or not recognized; the *insertions* or words that the system recognize and it never exist in the correct string; and the *substitutions* or words recognized as a different one compare with the correct string.

The WER error formulation is defined by the equation 2.4.

$$WER = \frac{100 * (D + I + S)}{N} \tag{2.4}$$

where N is the total number of words in the reference, D is the number of deletions, I is the number of insertions and S is the number of substitutions. The next example shows an hypothesized sentence compared with it reference and it WER calculation.

ref	SOMEDAY	YOU	WOULD	BE	BRING	ME	BACK	TO	***	YOU
hyp	***	***	LISTEN	TO	BRING	***	BACK	TO	THAT	GIRL
op	D	D	S	S	C	D	C	C	I	S

Where:
$D = 3$, $I = 1$, $S = 3$ and $N = 9$

$$WER = \frac{100 * (3 + 1 + 3)}{9} = 77.78\% \tag{2.5}$$

In this project, several experiments E will be averaged in order to obtain an overall results. The Formula 2.6 show the formulation of an average WER of E similar experiments.

$$WER = 100 * \frac{\sum_{e=1}^{E} D_e + I_e + S_e}{\sum_{e=1}^{E} N_e} \tag{2.6}$$

2.3.7 Existing state of the art for singing transcription

Taking advantage of the similarities between speech voice and singing voice, Mesaros [28] [29] [27] conducted several research. His studies were focused on systems the uses adapted HMM-GMM models trained from speech data. For conducting his research, a separation of the vocal from the audio signal using a *vocal separation* algorithm were used. This algorithm enhance the time-variant pitch and subtract the background model. In his studies, the isolation of the vocal leaded to an improvement of the performance of the models. Model using 39 MFCCs features and 39 mono-phones plus silence and short-pause with three HMMs states were constructed. It seems that use the *deltas* and *double deltas* coefficients in the training process were a necessary procedure in Mesaros case. This because the singing models were trained on speech data and adapted for sung models. Nevertheless, as was mentioned before, the variation on the formants between two phonemes in singing voice are very softly. Therefore, it expected that this coefficients in singing vectors features will not have a significant information neither a significant value variation.

The *Maximum linear likelihood regression* speaker adaptation technique was used by Mesaros in order to compensate the difference length on vowels in speech and singing voices. Nonetheless, it is not clear yet if with that technique he also compensate others singing characteristics like the *styles*. There are not discussion about that in Mesaros' work. For the *Language Model* a N-gram model where constructed using song lyrics with a goal of five thousands words vocabulary. The **HTK Toolkit** [54] was used in Mesaros' work.

The results of this work were a 57% accuracy using the system as a music retrieval and 24% correct lyrics transcription. As the author discussed, there are plenty of space for improve his work. This project starts from Mesaros' research and aims to investigate new approaches in order to address the problem.

2.4 Summary

Speech and singing are one of the most complex and important abilities of humankind. Moreover, seems that singing has been used for almost entire human history. Nevertheless, singing voice have some exclusive properties such style, pitch control, timbre and tempo. These properties define how a song is performed by an artist. Exist few studies using old HMM-GMM approaches for ASR that aims to extract the lyrics from an music audio-stream. The most important work in this topic using that approach was made by Mesaros [28] [29] [27] with an accuracy of 24% as a lyrics recognizer. The HMM-DNN approach are still not investigated in singing.

Chapter 3

Singing Corpus Design

Nowadays, exist several speech corpus that can be used for different SRE systems. Some of these database are for example, from broadband recordings; from spoken sequence of digits; from telephone speech; among many others. Most of these corpus are compiled by the LDC [25] which was founded in 1992 by several organizations in order to address the critical data shortage for language technologies. Nevertheless, an acoustic musical corpus it is missing in LDC and that motivate the design and construction of a new corpus called **Acoustic Cover Music version 1** or **ACOMUS1**. These chapter will start explaining the characteristics of the corpus following with the description of the design and the construction methodology.

3.1 ACOMUS Corpus version 1

ACOMUS1 it is the first version of a one channel, 16 Khz sample-rate corpus based on **Acoustic Covers** from popular music. The songs are interpreted by amateurs artists that using only one instrument, typically guitar, shared their recorders on YouTube. There are not other source rather than YouTube included in this version. Nevertheless, it is planning to investigate another sources like **SoundCloud** which is a platform where many amateurs artist distribute their songs. The corpus is composed by isolated segments of the spoken parts of the songs. The solo instruments segments are also individualised. The gender distribution only include male and female speakers, excluding children. This gender restriction is mainly because in YouTube only exist cover interpretations of songs made by adult people. Nevertheless, the current gender restriction applied in this first version of the corpus and may not be in possible future variations.

3.1.1 Database Collection

Besides YouTube, several academics projects that distribute some music collections were considered in the creation of ACOMUS1 [8, 17, 23, 45, 52, 2]. However, these projects are specifically oriented to musical researches rather than SRE. Therefore, the waveforms tend to be a polyphonic, multichannel and multi-speaker's original productions. In addition, these collections do not necessarily have lyrics which are mandatory for the annotation process. In contrast, on YouTube there are thousands of talented amateurs artist who share their acoustic covers interpretations to the general public. With all these artist is possible to obtain a big collection of songs that can be used in this corpus. In addition to YouTube advantages, obtain the lyrics of popular music for the annotation process is an straightforward task.

Audio Collection

For the first version of the database, a list of 240 YouTube's videos of acoustic covers collection from amateurs artist were selected. The database es divided by 200 guitar and 40 piano accompaniment instruments. The purpose of the piano covers samples is to perform some adaptation experiments. A complete list of these songs and its direct YouTube's link can be found in Appendix A.

In Table 3.1, the number of songs per guitar and piano per gender is detailed.

Table 3.1: Gender distribution of ACOMUS1 corpus

Gender	**Unique**	**Total**	**Instruments**	
	Singers	**Songs**	**Guitar**	**Piano**
Female	86	115	91	24
Male	89	125	109	16

This list contain 115 unique song separated by a balanced number of females and males singers. Also, repeated songs and artist are allowed in order to ensure different occurrences of sentences for adaptation techniques. The next Table 3.2 show the time size of the annotated songs in the database. The 100 guitar annotated songs size are 389 minutes nevertheless, the sung part is only the 60%. The other 40% are composed by introduction speech, solo instrument or discarded speech segments. The piano songs showed the same behaviour with a 58% of the total size as sung content.

Table 3.2: ACOMUS1 - Annotated Time Size

Instr.	**Annotated**	**Tot Time**	**Ann. Time**
Guitar	100	389 min.	233 min.
Piano	20	77 min.	49 min

Copyrights Consideration

Because the description of the data, the copyrights is an immediate concern that must be analysed. The **The UK Copyright Service** [48] outlines that the literary and the Musical works are protected. In ACOMUS1 corpus, the lyrics are tied to the literary copyright and the recordings sources are linked to the musical copyright. Nevertheless, as was mentioned before, the origin of the data is from a vast collection of covers from the YouTube service. Therefore, it can be assumed that the video covers already have the proper copyright permissions. In fact, in 2011 YouTube signed an agreement with two U.S. music publishers that allowed independent creators to keep up their covers, sharing the revenues with the composers [55]. Moreover, use the audio of these videos is assumed allowed because, will be only used for academic purpose and will never be distributed. Nor it has intension to generate a profit of any kind.

3.2 Corpus Structure

The final structure of ACOMUS1 corpus is a collection of isolated sung sentences. In order to separate the sentences, a complete annotation and posterior sentences extraction processes were designed and performed. The annotation of the database is a critical stage in the construction of the corpus because involve several manual processes that can lead to a different complications. One of the manual processes is to match correctly the lyrics with the song as is interpreted because the cover version do not necessarily follows the exact original lyrics. Another complication is to select which utterance are eligible for the final corpus. For example, some utterance are bad pronounced or incomplete because some artist errors or omissions and others, are totally unintelligible because the instrument domain the record over the voice.

In order to simplify the procedure, the process is separated into several sub-steps allowing to do an organize annotation of each song in a reasonable time. For the implementation of this stages, a re-utilization and adaptation of some tools from **CHiME** project were used and also, new tools for specific steps were created. The time necessary for the complete annotate process for a song is in average 30 minutes. For the first distribution of ACOMUS1 were annotated 100 guitar and 20 piano songs of the 240 total songs.

3.2.1 Annotation and Sentences Segmentation

As was explained above, the corpus was created from several acoustic cover songs extracted from YouTube. Nevertheless, the lyrics of a song have a particular grammatical structure. First of all, the speech in songs have a similar structure than a poem. However, it is not necessarily expected to have rimes but a similar number of syllable. These structure allows to treat each song as a set of smalls isolated utterance. These segmentation into sentences is made with the annotation procedure. The annotation process can be separated in two steps, the utterance individualisation and the creation of the corpus distribution files. The annotation results of ACOMUS1 is given in Appendix C.

Step 1: Utterance Individualization

In general, is expected that a cover interpretation of a song does not follow the exactly original lyrics content. Moreover, accidentally or purposely some covers tend to avoid or change some sentences and paragraphs modifying part of the structure of the original song. Because of these miss-match between the original song and the cover, it is necessary to match the lyric with the cover interpretation generating a new lyric with the specific cover characteristics.

Using the cover lyrics, each sentence must be identified and individualized and it is expected that each sentence in the lyrics corresponding to one individual utterance. Nevertheless, sometimes the particular interpretation of songs made unrecognisable when a sentences ends and the next start whenever the artist league the final vowel of the one sentence with the first vowel of the next sentence. When these link between sentences occurs, it is preferable to connect both sentences and transform its into one sentences re-modifying the cover lyrics. With this analysis, it is possible to finally establish the start and end point of each final annotated sentence. This is a manually and delicate step that must be performed with extremely care.

Step 2: Corpus Distribution Files

Using the segmentation of the previous step, a jsonification of the information it is performed. These json files allows to organize the info of the song detailing the start and end time of each sentence with the corresponding lyrics segment. This is the first step of the final construction of the distributed version of ACOMUS1 corpus.

Finally, a refine process of the start and stop time for each utterance and the possible modify of the sentence lyrics must be done. This process prevent overlapping of the singing speech between utterances, avoid missing words in the utterance and the miss-matching between the WAV raw and the lyrics. Moreover, the refine steps can ensure a best and precise individualization of each utterance for the final distribution. It is important to mention that the corpus distribution will never consider the final wav, it will only contain the final json files with the annotated information.

3.3 Summarize ACOMUS1 Characteristics

The Acoustic Cover Music corpus, is a corpus in development process that respond to the lack of an available music database. Moreover, ACOMUS1 is a corpus originated from acoustic version of music extracted from YouTube designed for been used in researches for academic purposes. The current stage of the construction of the corpus contain 200 songs of guitar plus voice covers and 40 piano plus voice covers. Nevertheless, only the half of the available songs had been annotated. It is expected that the size of the corpus grow in the next two years possible using additional sources such us SoundCloud service. The final corpus is conformed by a collection of several one channel isolated utterances that comprise only the sung segments of the songs sampling in 16 KHz.

Chapter 4

Evaluation Methodology

For the evaluation of this research, the common calculation of WER used in any SRE system will be implemented. This means that a count of the deletions, insertions and substitutions words will be performed. Nevertheless, this process must be performed in a fair way and a proper separation of training, development and testing data is needed. The training data is used for train the different HMM-GMM and HMM-DNN models and the development set is used in order to adjust the parameters looking for the best accuracy. Moreover, the test set it is only used at the end of the research in order evaluate the performance of the best model with an unseen set of data.

In order to organize the data, the characteristics of the corpus used were taken into account. ACOMUS1 corpus is a collection of isolated utterance from songs that can be sung by multiple singers and also, a singer can sung multiple songs. These distribution force to organize the data in a way that ensure that an specific song appears in one set whether is training, development or testing but not in another. Also, to ensure that a singer with all his interpreted songs appear in only one set. This organization gains importance when taking into account the objective of evaluate the performance of the system and determine if the ASR system can increase their robustness when is trained with artificial data augmentation.

In this chapter will be explained how each dataset were separated. Also, the different synthetic data options evaluated in this project will be presented.

4.1 Data Sets Separation

ACOMUS1 corpus is mainly oriented to acoustic cover versions with guitar as accompaniment instrument that currently having 100 annotated songs but also, contains 20 annotated pianos songs. Moreover, repeated songs and singers are allowed in the corpus therefore, in order to organize the database, it was divided into several rule based groups or **chunks** of 20 songs each.

- One song and all its different interpretations in the same chunk.
- All the songs form the same artist must be in the same chunk.
- All the songs in a chunk must contain the same accompaniment instrument.

With these rules and the current state of ACOMUS1 corpus, five guitar and one piano chunks were created. From the guitar chunks, arbitrarily the first four were selected for training and development sets and the fifth chunk and the piano one were separated as testing set. With the train and development chunks four models will be training intercalating the chunk used as development set and the ones used as training sets. This process have the intention to compensate the results from the small amount of data averaging the WER. The next Table 4.1 shows the number of utterance per model separated by training and development sets.

Table 4.1: Train and Development Size

Model	**Train Set**	**Development Set**	**Dev Weight**
Run 1	2034	556	21.5%
Run 2	1903	687	26.5%
Run 3	1846	744	28.7%
Run 4	1987	603	23.3%
Average	1942	648	25.0%

It is known that the size of the development set is at least twice as big as it should be. However, for the characteristics of the data it was not possible to create chunks smaller than 20 songs that follows the grouping rules. This distribution can be fixed whether increasing the annotated songs in ACOMUS1 corpus or using some augmentation in the training data for compensate the size of the corpus.

The testing guitar chunk does not have any special characteristic compared with the other previous four. However, as will never been used as training or development sets, this guitar chunk can be used for testing independently the final model selected. In addition, the purpose of the piano chunk is to perform the first analysis of the results of use a different background instrument than training. Table 4.2 shown the number of utterance per test sets.

Table 4.2: Guitar and Piano testing sets

Testing Set	**Utterance**
Guitar	637
Piano	770

4.2 Synthetic Augmentation

One of the objectives of the project is to evaluate if augmenting the database with synthetic modification can helps to increase the accuracy of the models. For this purpose, several synthetic data options were evaluated. Nevertheless, in order to keep a narrowed scope of the project, simple modifications of the original data will be tested.

Some of the modification that can be performed are to change the pitch, the tempo or add reverb effect. Nevertheless, only shift up and down the singer pitch and speed up and down the tempo will be the modification explored. The reverb effect was discarded in this project because require a previous audio analysis in order to determinate if the original audio have a natural reverb and in which grade. This analysis and implementation is to complex for a small time project like this. Nevertheless, the pitch and tempo modification is possible with less effort and as was explained in section 2.2, are important features in music.

With these modifications, the training sets can growth at least three times. In order to determinate which modification increase the accuracy several combination were used. Table 4.3 shown the size of the four combinations of modifications used. Firstly, in the first row it is shown the number of the original data per model, then in the second row it shown how the data grow when the pitch or the tempo modification is used. The third row shown the size using both modifications but without their combinations. Finally, the last row shown the size of the training data when tempo and pitch modifications and also its combinations are used.

Table 4.3: Training set size with Modifications

Modification	Models			
	Run 1	Run 2	Run 3	Run 4
With Original	2034	1903	1846	1987
With One Mod.	6102	5709	5538	5961
With Two Mod.	10170	9515	9230	9935
With combination	18306	17127	16614	17883

The next table, Table 4.4, shows how the weight of the development set over the augmented training data decrease in average from 25% to close to 3.5% which is more likely in any SRE.

Table 4.4: Weight Development set over augmented Training set

Modification	Models			
	Run 1	Run 2	Run 3	Run 4
With Original	21.5%	26.5%	28.7%	23.3%
With One Mod.	8.4%	10.7%	11.8%	9.2%
With Two Mod.	5.2%	6.7%	7.5%	5.7%
With combination	2.9%	3.9%	4.3%	3.3%

The details of how each modification was performed and the restrictions of its will be explained below.

4.2.1 Pitch Modification

For the pitch modification, different approaches were explored. The first approach was to try to calculate the *pitch* of the singer. Using that value, add and subtract a value that were inside the range of the gender of the singer. For calculate the pitch, several method were implemented such as; *Estimate frequency by counting zero crossings, Estimate frequency from peak of FFT, Estimate frequency using autocorrelation* [11] and *Estimate the tuning of an audio time series or spectrogram input* [24]. Moreover, an average of the results of different methods was tried. Nevertheless, even when these methods seems to be a accurate approach, none of them results in a value inside of the expected ranges. Because, the accompanier instrument add distortion at the results values obtained given the pitch value of the instrument plus the voice.

The final solution was try to find a gender independent average value that can be used to shift the pitch with the same constant value. In order to do this, two free tools were tested. The first tool was libROSA app [24] which offer a **pitch shift** function nevertheless, the acoustic results was a mix of the modified pitch with an undesirable reverb effect. The second option was to use SOX

[5] tool which allows to change the *pitch* in cents of a semitone. After several experimentation in males and females examples were found that increase and decrease the pitch in one tone results in a human like voice. Moreover, seems that when only one tone is of pitch is shifting with SOX tool, some vocal track length modification is added. Nevertheless, changes higher than one tone give in some cases a *chipmunk* voice.

4.2.2 Tempo Modification

For tempo modification were also tested several methods in order to obtain results into the ranges of real music. The initial method of this modification was to try calculate the original tempo of the song and, using rule based system, increase or decrease the speed of the song. Nevertheless, even when libraries like libROSA were used for calculate the **beats** of the song, the result tent to represent the instruments tempo rather than the voice. Therefore, some song had a completely mismatch between the beats calculated and the speed of the voice.

The solution was to find a subjective fixed value where the *tempo* sounds "humanly" sung. The final value for the *tempo* modification was obtained after experimenting several values of velocity looking for a significant differentiation from the original song but inside a humanly music speed range. The final parameters obtained was a variation of a ±15% of the original song. In order to maintain the simplicity and the coherence of the project this modification was also made using SOX tool. The main advantage of use SOX to perform tempo modification is that the tempo function do not add a pitch variation. Therefore, the result obtained are more desirable for the experiments that will be implemented allowing test only the results of add speed variation rather speed mixed with pitch variation.

Chapter 5

System Development

As was mentioned before, the database corpus used in this research is ACOMUS1 which is constructed from several isolated sentences from covers songs. The Language Model necessary for this research must be related with these corpus therefore, a lyrics based LM from real music was be used. Nevertheless, this kind of LM does not currently exist consequently, the design and construction of a new LM is necessary. In this chapter, different aspects of the designed LM and of the AM selected will be discussed.

5.1 Language Model

The design of the required LM for this project was directly related with the ACOMUS1 database. This relation was mainly with the sentences structure and the vocabulary used in a song. The structure of a song it is similar than a poem but, with a less elaborated vocabulary and not necessarily using rimes. However, like a poem the message of a song is given with imagery rather than explicit. This message is usually expressed using several sentences that are not connected with any linker. As an example, the first paragraph of the lyrics of the song **More Than Words** from the band **Extreme** is expressed below.

> Saying I love you
> Is not the words I want to hear from you
> It's not that I want you
> Not to say, but if you only knew
> How easy it would be to show me how you feel
> More than words is all you have to do to make it real
> Then you wouldn't have to say that you love me
> 'Cause I'd already know

It is possible to see in the previous song's extract that the paragraph have a clear message but it structure is similar than a poem where each sentence can be treated independently.

5.1.1 Language Model Construction

In order to elaborate a LM with similar vocabulary and grammar structure than the database, the LM was based on the lyrics from the discography of the 157 original artist contained in ACOMUS1 corpus. In addition, a complementary list of 39 popular artists were appended to increase the final size of the LM corpus. The total number of songs obtained from these 196 artist were 25,916. The complete list of the artists used for the LM model can be found in Appendix B.

The main target of the design of the LM was the creation of a fair LM. This means that the LM should not assign high probabilities to a song depended sentences. But also, should include the poem structure where the sentences are not always grammatically correct. In the process of the construction of the LM first, one global corpus that group all the collected lyrics was created. Next, over this great corpus, several rules were applied in order to exclude the sentences that could add noise or assign an un-fair probabilities to specific sequence of words.

LM: Song Level Rules

Two song levels rules were considered in order to create a LM independent of the database song. The first rule was to exclude the lyrics of the songs contained in ACOMUS1 in order to avoid to assign a valid likelihood to a song depended sentences. For example, the sentence **A Mosquito My Libido** from the song **Smells Like Teen Spirit** from **Nirvana** is a very uncommon sentence exclusive from these song. In contrast, the sentence **I LOVE YOU** can be founded in several songs from different artist with different styles like the punk-rock band **The Ramones** with the song **Baby I Love You** and the rock band **The Beatles** with the pop song **P.S. I Love You**.

Secondly, songs with the same name of the ones in ACOMUS1 were also excluded in order to filter all possible covers versions from different artists. For instance, the same song **Baby I Love You** from the **The Ramones** it is in fact a cover version of the **The Ronettes'** version. Nevertheless, this *song name* exclusion also consider songs that are completely different but share the same name. This is evident in the case of the song **How Deep Is You Love** where even when the **Bee Gees** and **Calvin Harris** have a song with that name, its lyrics are completely different and therefore, are complete different songs.

LM: Sentences Level Rules

Seven sentences level rules were created in order to normalize the corpus content.

1. A words normalization in the corpus was performed. These normalization involve to transform cardinal numbers into words and also, clean repeated letters in a word. For example, it is common to find the word **yeah** written with more than one **e** when was intended to express an extension of the word.

2. Using CMUdict lexicon [7], all the OOV words were founded and marked with an <UNK> tag.

3. All the sentences with more than two OOV words were excluded. This exclusion allows to eliminate the sentences written in different language than English. For Example, in the song **Spanish Bomb** from the **The Clash** exist several sentences singed in Spanish even when the song is in English.

4. Exclude the metadata sentences that can be found in the content of a lyrics. The most typical example is the *chorus* tag that indicate when the chorus of the song start. But in general, all the sentences that starts and ends with brackets are consider metadata in the lyrics.

5. Only four repetitions of a sentence in a song were allowed. This rule was very important in order to keep a fair model. For instance, in the song **All You Need Is Love** from the **The Beatles**, the sentence **All you need is love** is repeated more than 25 times but for the LM porpoise only four repetition were used.

6. Misspelling normalization were also applied. In general, the lyrics that can be found on **The Internet** are wrote manually by fans therefore, find some misspellings were expected. However, mainly these misspellings were abbreviations like the word **'CAUSE** and **CAUSE'**.

At the end of this normalization process, sentences constructed only with words contained in the final AM lexicon were kept. The vocabulary used will be explained in Section 5.2.

5.1.2 Creating the lyrics discography

As was mentioned before, the corpus was constructed using a collection of lyrics from the discography of popular artists. This collection of lyrics was obtained from the website *http://lyrics.wikia.com*. Each artist was saved into independent directories that contain a file text per each song. Save the lyrics in this way allows to work with the artist and his song independently making it easer to manage, modify, add or delete some specific song.

The final corpus created for this research have $509,844$ lines with a total of $3,226,932$ words. The words count does not taking into account the start (<s>) and end (</s>)sentence token. This corpus have in average six words per sentences demonstrating that the structure of a common lyrics is a set of small sentences.

5.2 Acoustic Model

The vocabulary for the acoustic model was based in the **CMUdict** lexicon dictionary [7] but, in order to keep a simple model only five thousands words (5K) were selected [35, 20].

The 5K vocabulary was constructed ensuring to keep most of the annotated utterance. In order to do this firstly, all the intersected words between the lyrics of the database and the CMUdict were used. With the 240 song contained in ACOMUS1, the vocabulary obtained were around 3,500 words. Therefore, in order to complete the whole 5K vocabulary a random selection of words from the LM corpus was made without altering the original distribution of its. In order to keep the random selection of the rest of the words robust, a fixed random seed was used. Therefore, the only way that this selection would changes is if the database is changed, is increased or if the LM corpus is modified somehow.

As is expected, exist several words that appears in the lyrics but are not in the CMUdict resulting in a collection of missing words. The analysis of these words will be explained below.

5.2.1 Words analysis in CMUdict

At the moment this research was made, the CMUdict vocabulary had 39 phonemes and 133,031 different words. The phonemes used in CMUdict can represent correctly the English language nevertheless, as is expected this vocabulary does not contain all the possible words that can be found in the lyrics collection. The different possible words that can founded in the lyrics are virtually uncontrollable. This because of the existence of; missing words in the base lexicon, invented words and different slang

In order to decide what would be the best procedure in this matter, a detailed analysis of the OOV words was made.

Missing, Invented and Slang Words

Exist several words in the lyrics collection that could be used in the project's vocabulary but, for some reasons are not possible to find in the CMUdict. The first case are valid words that are not contained into CMUdict dictionary. For example, the term *Wonderwall* is a word that have a clear have a meaning and is used by **OASIS** as on one of their songs but, it is not part of the CMUdict dictionary. After an analysis of the LM corpus, it was founded only few cases of these kind of words. Moreover, this words tend to be song dependent and not necessarily common.

Secondly, it was expected to find few cases of invented words in the LM corpus. Nevertheless, after a complete analysis was discovered that no invented words appears in the LM corpus. But, this not necessarily means that if the LM is grown with new artist these words could not appear.

And finally, slang words were also expected and founded in the LM corpus. However, only was with the rap artist **50 Cents** and with no other artist. These may confirm that the vocabulary can be connected with the music style.

In these three cases the decision was not to add the words into the vocabulary because it could demand phoneticist support and only affect few samples in the whole collection.

Chapter 6

Experimental Results

The target of this research is to attempt to determine if use augmenting training data can increase the performance of ASR models on music corpus and also, evaluate the effect in the performance using a DNN approach. In order to answer these questions, a set of experiments on ACOMUS1 corpus were performed, first, to determine the optimal N-gram size to use and second, to evaluate different configurations of augmented training data on GMM and DNN acoustic densities models.

This Chapter is divided in six Section that will started explaining common methodologies and terminologies of the experiments performed. Following by several experiments organized by the objectives of each of them. At the end of this Chapter a summarize of the experiments can be found.

In detail, this Chapter is divided as:

Section 6.1: Section that describe common terminology and approaches used in the experiments.

Section 6.2: Section that explain the N-gram selection experiments.

Section 6.3: In this Section will be detailed the baseline construction for the following experiments.

Section 6.4: The experiments in augmented training data will be shown in this Section, detailing the results in pitch and tempo modifications explained in Section 4.2.

Section 6.5: The final evaluation of the guitar and piano test sets decoded using the best models obtained in the previous experiments will be show in this Section.

Section 6.6: Section that contain a final summarization of the experiments detailed above.

6.1 GMM & DNN Acoustic Densities Models

For each experiment in this project, the same sequence, or part of it, of training process were performed. This experiment were a progressive training of acoustic densities models that in the case of the GMM models, started with a mono-phone training and ended with a speaker adapted tri-phone. For the neural networks a DNN and a DNN sMBR models were trained.

For the technical implementation of this project see Appendix C.

For the GMM models, the four run1 to run4 experiments explained in Chapter 4 were executed and averaged for consolidates results. The models trained, in the execution order, were:

1. **Mono:** Mono-phone system trained using MFCC features plus delta + delta-delta.
2. **Tri1:** Tri-phone system trained using MFCC plus delta + delta-delta.
3. **Tri2b:** Tri-phone system trained using LDA+MLLT features.
4. **Tri3b:** Tri-phone system trained using SAT over LDA+MLLT features [37].

Because of the high executing time needed for the DNN models, only run1 and run2 experiment were executed and averaged. The DNN models trained were:

1. **DNN:** DNN model trained using fMLLR features.
2. **DNN sMBR:** DNN model with sMBR.

6.2 Determining the N-gram size

The LM used in this project is based on lyrics from popular music that contain six words length in average per sentences (Section 5.1.2). Therefore, it was expected that the optimal N-gram model size should be between bigram and trigram models This because, a unigram model is more proper size for SRE systems based on isolated words like a sequence of digits where no relation between words in a sentence existed. In contrast, a higher N-gram size model is more likely for systems based on extended speech structure like news broadcasting or telephone conversations.

In order to determine the best N-gram size for the LM corpus, a set of four experiments on GMM's models were performed. The procedure in these experiments was to run the run1 experiment described in Section 4.1 with an incremental N-gram size starting from unigram and increasing to 4-gram. These experiments were executed two times in order to corroborate the results.

It was shown that the best result was obtained using a bigram model with an WER of 90.49%. Moreover, as was expected the unigram model had the lower performance. In addition, trigram and 4-gram models had a slightly better performance than unigram. Nonetheless, all the results have a high WER over 90%. In Table 6.1, the results of these four experiments are summarize.

Table 6.1: Bigram models shown the lower WER with 90.49%

Model	Unigram	Bigram	Trigram	4-gram
Mono	93.58%	**92.01%**	92.68%	93.50%
Tri1	92.94%	**91.42%**	92.09%	93.11%
Tri2B	94.03%	**92.43%**	93.02%	93.20%
Tri3B	92.66%	**90.49%**	90.74%	91.08%

As the objective of these experiment was merely to obtain the optimal LM size a major analysis of the reason for the high WER was discarded. With the results obtained, a **bigram** LM models was selected for the following experiments.

6.3 Baseline Results

Before to analyse the effect of modifications training data, elaborate a baseline results were necessary. For its construction, a naive approach that train the experiments run1 to run4 on the original data were used. The GMM's results showed that from Mono to Tri3B models the insertions decreased in 27% and the substituted words decreased in a 11%. In contrast, the deleted word increased in a 15%. This results gave an improvement of performance with only one percent from 93.77% to 92.70% WER. Moreover, for the DNN models, the average results were similar than GMM adding no significant improvement. In Table 6.2 the details of the evolution of the inserted, deleted and substituted words are summarized.

Table 6.2: Average results of baseline experiments separated by GMM and DNN results. See text for details.

Model	**Total Words**	**Insertion**	**Deletion**	**Substitution**	**Average WER**
Mono	17109	429	6312	9302	93.77%
Tri1		373	7215	8476	93.90%
Tri2B		353	7399	8301	93.82%
Tri3B		315	7269	8275	92.70%
DNN	8219	234	2954	4410	92.44%
DNN SMBR		194	3054	4249	91.22%

Even when the WER had a slightly overall decrease, the high number of deletion words can be an indicator of the necessity of increase the database corpus in order to obtain more examples of some words and may also indicate the necessity of to grow the LM up. In Table 6.3 an example that lustrate how some substitution errors are in fact, a critical error because the substitution are not with homo-phones. Instead, Table 6.4 show an example where 75% of the words in the sentence were deleted. Nevertheless, in the second example, an alignment error can also be notice because the original **WOULD LIKE** it is phonetically close to the transcription **WHO LIKES**. However, for the system **LIKE** aligned with **LIKES** is equally wrong than **SAY** aligned with **LIKES** and the calculation of the likelihood of the best path align the sentence in that way.

Table 6.3: Substitution error in a sentence from the song from Adele, *Send My Love*.

ref	SEND	MY	LOVE	TO	YOUR	NEW	LOVER
hyp	***	SOMEONE	TO	FEEL	LIKE	A	FIGHT
op	D	S	S	S	S	S	S

Table 6.4: Deletion error in a sentence from the song from Oasis, *Wonderwall*.

ref	THERE	ARE	MANY	THINGS	THAT	I	WOULD	LIKE	TO	SAY	TO	YOU
hyp	***	***	***	***	***	***	***	***	WHO	LIKES	ME	NOW
op	D	D	D	D	D	D	D	D	S	S	S	S

The detailed results per experiment can be founded in Appendix D.

6.4 Results with Augmented Training Data

These experiments were divided in four different combinations of modifications for the augmentation of the training data. As equal than the baseline experiments, a codification was used in order to group and average the final results. The codename for each experiments are:

- RunX.1: Experiment using pitch modification.
- RunX.2: Identify models with tempo augmentation.
- RunX.3: Identify models with pitch and tempo augmentation without its combinations.
- RunX.4: Identify models with pitch and tempo augmentation with its combinations.

These four augmented training data experiments were applied over run1 to run4 experiments, augmenting the train set of each experiment and averaging their results.

Table 6.5: Pitch augmentation gave a slightly improvement in the results with WER 89.61%.

Model	**Average WER**				
	Baseline	runX.1	runX.2	runX.3	runX.4
Mono	92.01%	**93.78%**	93.71%	93.16%	94.09%
Tri1	91.42%	**92.93%**	94.38%	93.38%	93.40%
Tri2B	92.43%	**93.30%**	94.23%	93.49%	93.52%
Tri3B	90.49%	**89.61%**	93.07%	91.70%	91.62%

Table 6.5 shows that the tempo modification have no positive effect in the average WER increasing the error in 2.5%. As the tempo modification just repeat the same features but stretched and shortened in time, there are no new information for the system.

The combined models runX.3 and runX.4 got also higher error rate than baseline, this results were expected after the results with the tempo modification. In contrast, the pitch modification is the only augmentation technique that have an effect decreasing the WER. As it is notice in the table, the improvement occur when a speaker adaptation technique is used. The pitch modification was added without change the speaker assigned to that waveform, this may indicate that the system assign a higher pitch range for each speaker allowing it a better transcription. Nevertheless, a decrease of 3% from 92.70% to 89.61% of WER it is too small to be consider an improved solution.

For the DNN's experiments, only run1 and run2 experiments with pitch augmented training data were tested. The overall results of these experiments are shown in Table 6.6. These DNN approaches help to decrease the WER in 1% from 89.61% obtained with Tri3B to 88.33% with DNN sMBR.

Table 6.6: DNN sMBR model had a 88.33% WER in development data showing a decreasing of 4% with the baseline.

Model	Average WER
DNN	90.87%
DNN sMBR	88.33%

6.4.1 Pitch Augmentation as New Same Gender Speaker

Additional to the previous experiments, a variation of the pitch training augmentation experiments *runX.1* were tested. In this experiments, each new modified audio was labelled as a new speaker of the same gender than the original increasing the training dataset by ”new” speakers three times. The main variation expected with this approach was with the speaker adaptation in Tri3B models. Nevertheless, this approach helps to increase the WER as well as the previous experiments. This results can endorse the assumption that in this system the Tri3B model can improve the performance when the pitch range per speaker is higher than the normal range. The overall results from the HMM-GMM models for these experiments are showed in Table 6.7.

Table 6.7: Augmented Data Results

Model	**Average WER**
	RunX.1S
Mono	93.38%
Tri1	93.44%
Tri2B	93.27%
Tri3B	92.00%

6.5 Test Data Sets Results

The best model obtained was a bigram LM, using pitch augmented training data. The WER obtaining was 89.61% with Tri3B model and 88.33% with DNN sMBR model. These characteristics represent the experiments runX.1 described in Section 6.4. Using this model, both test sets, with guitar and with piano background were decoded. The final results are expressed in Table 6.8

Table 6.8: Guitar and Piano Test Set Results.

Model	**Background**	
	Guitar	**Piano**
Tri3B	90.80%	86.87%
DNN MPE	89.88%	84.59%

As it is expected, the guitar test set obtained a slightly higher WER than the results with the development set. Nevertheless, when the model was used for decode the piano test set, the WER error decreased in the GMM and DNN models. The lower error in this case can be explained firstly, by the difference of frequencies ranges of the instruments (Section 2.2.1). This because, as the piano have a higher range than the guitar it is expected that more notes are played outside the range of the singing voice. Secondly, this difference can also be explained by the difference position of the microphone when a piano song is recorder and when a guitar song is recorder. For example, when a piano song is recorder, usually the microphone is positioned over the piano closer to the singer capturing more power from the voice than the instrument. However, when a guitar song is recorder the microphone is positioned between the instrument and the singer capturing more energy from the instrument.

6.6 Experiments Summarize

In overall, the baseline results got a 92.70% using a triphone+SAT GMM model and an slightly improve to 91.22% using DNN sMBR model. Moreover, the synthetic augmented training data experiments performed in order to improve the results showed that the pitch modification had the best performance decreasing the error to 89.61% on GMM models and 88.33% on DNN model. This improve can be notice when a speaker adaptation technique is applied. It is claim that as the pitch modification applied followed a fixed value it created a higher range of singer pitch helping the SAT technique to improve the adaptation. Nevertheless, the results are far from being a positive results. Therefore, it is argue that the synthetic modifications applied do not decrease the WER in a small database like the used in this project.

Chapter 7

Conclusions

This project was undertaken to design and construct an Acoustic Cover Music corpus and evaluate if to use synthetic training augmented data as well as HMM-DNN models can improve the performance of a SRE system on this corpus. For this purpose, experiments with the original data where performed in order to elaborate a research baseline and also, pitch and tempo modifications were applied in order to performed the augmented training data experiments. The major limitation of this study is the small corpus used that was completely created from scratch by the author. This study has identified that pitch modification technique was the only one that showed a positive variation decreasing the WER from 90.49% baseline to 89.61% using GMM models and to 88.33% using DNN sMBR models. This improvement may indicate that to increase the speaker pitch range helps the efficiency of the SAT technique. However, these variation are to small and the error range is still to high. The results of this study indicate that to use pitch or tempo modifications to increase the training dataset is a technique that fail in the aim to decrease the WER. In addition, in this conditions DNN models also failed improving the performance of the models obtaining a very small decrease of the WER. This is the first study to investigate the effects of augmented training data and DNN models in a novel musical corpus creating a baseline for future researches or challenges.

7.1 Future Works

There are several modifications that can be applied in different stages of this project. For the previous stages, a vocal separation techniques can be explored in order to determine its effectiveness and elaborate a parallel control model. For the LM model, the construction of a plagiarism system may be interesting to be explored in order to create a more fairly model. Elaborate a more sophisticate methodology for the synthetic augmented training data for the pitch modification and also, explore others modifications likes reverb effect.

7.1.1 ACOMUS1

ACOMUS1 is a newer corpus that currently is in a increasing stage therefore, some improvements can be applied One improvement that may be necessary is to add other possible sources rather than YouTube in order to expand the samples options. Also, increase the size of the guitar and piano samples to more than a thousand songs will be desirable. Finally, increase the background instruments from guitar and piano to others like violin or violoncello.

Bibliography

[1] BARTHOLOMEW, W. T. A Physical Definition of ”Good Voice-Quality” in the Male Voice. *The Journal of the Acoustical Society of America 5*, 3 (2005), 224.

[2] BITTNER, R., SALAMON, J., TIERNEY, M., MAUCH, M., CANNAM, C., AND BELLO, J. P. MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *5th International Society for Music Information Retrieval Conference* (Taipei, Taiwan, 2014).

[3] BOLL, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing 27*, 2 (1979), 113–120.

[4] CAMBRIDGE UNIVERSITY ENGINEERING DEPARTMENT. HTK Speech Recognition Toolkit.

[5] CBAGWELL, ROBS, AND UKLAUER. SoX - Sound eXchange.

[6] CHAVAN, K. Speech Recognition in Noisy Environment , Issues and Challenges : A Review.

[7] CMU. The CMU Pronouncing Dictionary.

[8] DE MAN, B., MORA-MCGINITY, M., FAZEK, G., AND REISS, J. D. The Open Multitrack Testbed. In *137th Convention of the Audio Engineering Society* (2014).

[9] EL-FATTAH, M. A. A., DESSOUKY, M. I., DIAB, S. M., AND EL-SAMIE, F. E. A. Adaptive Wiener Filtering Approach for Speech Enhancement. *Ubiquitous Computing and Communication Journal 3*, 2 (2010), 23–31.

[10] ELLIS, D. P. W. Extracting Information From Music Audio. *Communications of the ACM 49*, 8 (2006).

[11] ENDOLITH. Frequency estimation methods in Python.

[12] EPHRAIM, Y., AND TREES, H. L. V. A Signal Subspace Approach for Speech Enhancement. *IEEE Transactions on Speech and Audio Processing 3* (1995), 804–807.

[13] GEIRHOFER, S. Feature reduction with linear discriminant analysis and its performance on phoneme recognition. Tech. rep., University of Illinois at Urbana-Champaign, 2004.

[14] GOTO, M., ITOU, K., AND KITAYAMA, K. Speech-recognition interfaces for music information retrieval. *Music Information Retrieval* (2004).

[15] HAIN, T. Speech Technology - Handout 3a - Neural Network in Speech Technology, 2016.

[16] HINTON, G. A Practical Guide to Training Restricted Boltzmann Machines A Practical Guide to Training Restricted Boltzmann Machines. *Computer 9*, 3 (2010), 1.

[17] JANER, J. Audio Signal Separation | Music Technology Group, 2013.

[18] JURAFSKY, D., AND MARTIN, J. *Speech and Language Processing*, second edi ed. Parson, 2000.

[19] KAN, M. Y., WANG, Y., ISKANDAR, D., NWE, T. L., AND SHENOY, A. LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech and Language Processing 16*, 2 (2008), 338–349.

[20] KING, S., BARTELS, C., AND BILMES, J. SVitchboard 1: Small Vocabulary Tasks from Switchboard 1. In *INTERSPEECH-2005* (Lisbon, Portugal, 2005), pp. 3385–3388.

[21] KOB, M., HENRICH, N., HERZEL, H., HOWARD, D., TOKUDA, I. T., AND WOLFE, J. Analysing and understanding the singing voice: Recent progress and open questions. *Current Bioinformatics 6* (2011), 362–374.

[22] LATHAM, A. *The oxford companion to music.* USA: Oxford University Press., 2002.

[23] LERCH, A. Audio Content Analysis - Datasets, 2015.

[24] LIBROSA. librosa.feature.delta — librosa 0.4.2 documentation.

[25] LINGUISTICS DATA CONSORTIUM. Linguistic Data Consortium.

[26] LUNDY, D. S., ROY, S., CASIANO, R. R., XUE, J. W., AND EVANS, J. Acoustic analysis of the singing and speaking voice in singing students. *Journal of voice : official journal of the Voice Foundation 14*, 4 (2000), 490–493.

[27] MESAROS, A. Singing voice identification and lyrics transcription for music information retrieval. *2013 7th Conference on Speech Technology and Human - Computer Dialogue, SpeD* (2013), 1–10.

[28] MESAROS, A., AND VIRTANEN, T. Automatic Recognition of Lyrics in Singing. *EURASIP Journal on Audio, Speech, and Music Processing 2010*, 4 (2010), 11.

[29] MESAROS, A., AND VIRTANEN, T. Automatic Understanding of Lyrics From Singing. *Korkeakoulunkatu 1, 33720, TAMPERE* (2011), 1–6.

[30] MESAROS, A., VIRTANEN, T., AND KLAPURI, A. Singer identification in polyphonic music using vocal separation and pattern recognition methods. *In Proc. 7th International Conference on Music Information Retrieval* (2007), 375–378.

[31] MOORE, R. Speech Processing - Lecture 03 - Speaking, 2015.

[32] MOORE, R. Speech Processing - Lecture 05 - The Nature of Speech., 2015.

[33] MOORE, R. Speech Processing - Lecture 19 - Cepstral Analizys, 2015.

[34] Omori, K., Kacker, A., Carroll, L. M., Riley, W. D., and Blaugrund, S. M. Singing power ratio: Quantitative evaluation of singing voice quality. *Journal of Voice 10*, 3 (1996), 228–235.

[35] Paul, D. B., and Baker, J. M. The Design for the Wall Street Journal-based CSR Corpus. In *Proceedings of the Workshop on Speech and Natural Language* (Stroudsburg, PA, USA, 1992), HLT '91, Association for Computational Linguistics, pp. 357–362.

[36] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., and Others. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (2011), no. EPFL-CONF-192584, IEEE Signal Processing Society.

[37] Povey, D., Kuo, H. K. J., and Soltau, H. Fast speaker adaptive training for speech recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2008), 1245–1248.

[38] Psutka, J. V. *Benefit of Maximum Likelihood Linear Transform (MLLT) Used at Different Levels of Covariance Matrices Clustering in ASR Systems.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 431–438.

[39] Quinn, S., and Watt, R. The perception of tempo in music. *Perception 35*, 2 (2006), 267–280.

[40] Raj, B., Parikh, V. N., and Stern, R. M. The effects of background music on speech recognition accuracy. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (Munich, 1997), IEEE, pp. 851–854 vol.2.

[41] Rechel, L. M. The effect of singing tempo during specific song acquisition of preschool aged children, 2013.

[42] Rosenblatt, F. Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms. *Archives of General Psychiatry 7* (1962), 218–219.

[43] Rosillo Gil, V. *Automatic Speech Recognition with Kaldi toolkit.* PhD thesis, 2016.

[44] Rumelhart, D. E., and Zipser, D. Feature discovery by competitive learning. *Cognitive Science 9*, 1 (1985), 75–112.

[45] SiSEC 2013. SISEC 2013 : HomePage, 2013.

[46] Sundberg, J. *The Science of the Singing Voice.* Northern Illinois University Press, Dekalb, Illinois, 1987.

[47] Sundberg, J. Vocal tract resonance. *Vocal Health and Pedagogy, Volume: Science and Assessment 1* (2006), 103.

[48] The UK Copyright Service. Fact sheet P-01: UK Copyright Law, 2015.

[49] Tokuda, K. Drawing Figures of Hidden Markov Models in A T Xwith PSTricks. 1–3.

[50] Tsai, W.-H., Liao, S.-J., and Lai, C. Automatic Identification of Simultaneous Singers in Duet Recordings. In *9th International Conference of Music Information Retrieval* (Philadelphia, 2008), pp. 115–120.

[51] Tsai, W. H., and Wang, H. M. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 1 (2006), 330–341.

[52] Vincent, E., Gribonval, R., and Plumbley, M. D. BSS_ORACLE: A toolbox to compute oracle estimators for source separation, 2007.

[53] Weiss, R. J., and Ellis, D. P. W. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech and Language 24*, 1 (2010), 16–29.

[54] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. The HTK Book (for HTK Version 3.4). *Cambridge University Engineering Department*, July 2000 (2009), 384.

[55] YouTube. Official YouTube Blog: Creating new opportunities for publishers and songwriters, 2011.

[56] ZyTrax, I. Tech Stuff - Frequency Ranges.

Appendices

Appendix A

Sources for ACOMUS1

The following table contain the complete list of videos links used in this project. For access to each link use: **https://youtu.be/{YouTube Link}**

Table A.1: List of guitar songs

Cover Artist	**YouTube Link**	**Original Artist**	**Song Name**
220volt74z	eOuQ-xVEqK8	Scorpions	Send me an Angel
ACMusic7	QvyglEro34s	Foo Fighters	Learn To Fly
Ady Suleiman	2JK9G0vdbro	Miguel	Quickie
Alan Robinson	KQyMoky6oK4	Leonard Cohen	Hallelujah
Alan Robinson	JwGcobw6mD0	The Beatles	From Me To You
Alan Robinson	uRGz08FKaiE	The Beatles	Let It Be
Alan Robinson	oEtoA–yyfA	The Ramones	Baby I Love You
Alex Aiono	MAh_OXCj0IA	James Bay	Let It Go
Alex Francis	PFLrAUGGWFU	Michael Jackson	Love Never Felt So Good
Alex Hamel	NS4SrfV9PKQ	Elton John	Your Song
Alex Hamel	Whvwh_uqH5s	Eric Clapton	Tears In Heaven
Alex Hamel	k1FOnYnW7ok	Radiohead	Creep
Alex Hamel	0J8SoKlyJHs	The Beatles	Come Together
Alex Hamel	Jt5Njb0bY4g	The Doors	Light My Fire
Alexandra Burke	81cjT12FRaw	Janet Jackson	Let's Wait Awhile
Amanda Law	8BpXAA1bqUk	Justin Bieber	What Do You Mean
Amber Whitworth	ucmZjeJ7C8Y	Beyonce	Pretty Hurts
Andreas Moe	gFOwWCpsgGw	A Great Big World	Say Something
Andrew Garcia	4YS2qfVt4UQ	MAROON 5	Sunday Morning
Andrew Garcia	d9Sipy8t0Il	Oasis	Wonderwall
Andrew Garcia	-4Yx_GYcYdQ	Rihanna	Work
Arlene Zelina	n-PPzymcxzA	Pharrell Williams	Happy

Awake! Awake!	k2edmOOC6G8	Coldplay	Midnight
Awake! Awake!	5JuUMoZKFsk	Queen	Save Me
Awake! Awake!	10dp0NZtyf0	U2	Ordinary Love
Ayelle	q0-_3fZBGuQ	Sia	Elastic Heart
BaneSIlvermoon	GcBOCelWoDE	Pearl Jam	Black
Beth Sherburn	OT4lg0ZzQSM	Michael Jackson	Love Never Felt So Good
Birdy	Z8Qnui5DWg4	Birdy	Tee Shirt
Boyce Avenue	yhAIVxcE3QA	Eric Clapton	Tears In Heaven
Boyce Avenue	Z1IpZRe7-R8	Foo Fighters	Everlong
Boyce Avenue	0SNCkrC2UCU	John Mayer	The Age of Worry
Boyce Avenue	FQNN5kCbQcQ	Oasis	Wonderwall
Braeden Counts	OPK3jt5WAAA	Florida Georgia Line	Anything Goes
Brian Mikula	ekv68VNC4CI	Elton John	Tiny Dancer
Brian Mikula	3EvCSZglEWI	Foo Fighters	My Hero
Brian Mikula	qX7zymWSw70	Pink Floyd	Comfortably Numb
Brian Mikula	JLcl2698NaY	Tonic	If You Could Only See
Brodie Kelly	mwmj7Hsfqsk	Cody Simpson	Free
Cammie Lester	theFtkdm5Ps	Chris Stapleton	Tennessee Whiskey
Charlie Brown	Ba34X-0REUo	Miguel	Adorn
Chromatone	3TXKLJZjpSc	Rita Ora	I Will Never Let You Down
Clara Bond	7Q6q_nLOgPk	Shawn Mendes	Stitches
Conor Coughlan	seHEuq-hipg	Jamie Lawson	Wasn't Expecting That
Craig	_UWDuk5j92w	Bouncing Souls	Kids and Heroes
Daniela Andrade	fzxag7U3Snk	Gnarls Barkley	Crazy
Daniela Andrade	LkxSU__Bi3w	Nirvana	Smells Like Teen Spirit
Daniela Andrade	6Zex2x-6ePc	Pixies	Where Is My Mind?
Daniela Andrade	uQzm3GdqfMM	Rihanna	Higher
Danny McEvoy	eRJmAPuJOBs	Peter Cetera	Glory of Love
De La Rose	Nwa2Hvg9o-U	Lianne La Havas	Ghost
Delaire	oALNHO-qQCA	Sam Smith	I'm Not The Only One
Dennis Baumann	waBMrTY7xVA	Bad Religion	Sorrow
Ebony Day	ksNdLcWIG2w	Calvin Harris	How Deep Is Your Love
Ebony Day	6wnNT1WSGgY	Justin Bieber	Love Yourself
Ebony Day	4F818SITrxk	Katy Perry	Roar
Ed Sheeran	AnFsq1Jx-yQ	Lorde	Royals
Effie	2jawiv5UmRc	Kwabs	Wrong Or Right
ELIZA	3hPZafPMpeI	Alessia Cara	Here
Eliza Shaddad	CE-3NVMn2iw	Kiesza	Hideaway

Elliott Pritchard	MwFmkDI3bI0	Rihanna	FourFive Seconds
EMG	75Q05_I11wA	Adele	Someone Like You
EMG	uY6GPeeZ9aQ	Al Green	Let’s Stay Together
EMG	EPCNnL2TkDc	Lionel Richie	Hello
EMG	iI5HLPnVGt0	The Police	Every Breath You Take
EMG	2xUHkiANYts	Tori Amos	Winter
Emily Davis	-2rMeUkvpyI	Bad Religion	You
Eskimos	x0hoWvBZMfk	Katy B	Crying For No Reason
Etham	U9yZ9FDbgYk	Adele	Hello
Etham Basden	fpE_4RRc8rs	Bruno Mars	Gorilla
Fabio Righi	0K0QIaVF4gA	Rolling Stones	Paint It Black
Florence Pardoe	3AF74JgDUS8	John Lennon	Jealous Guy
Flynn	W9IyAqoZnB0	Jennifer Hudson	Trouble
Fraser Churchill	8tzHN2mct9A	MAGIC!	Rude
Fré Monti	kuaFUflWwdU	Seal	Kiss From A Rose
Grace Weber	NJA9Io_YbeE	Stevie Wonder	Isn’t She Lovely
Gustavo Trebien	d-Zd6q7cdI4	Creed	My Sacrifice
Gustavo Trebien	wRzA0gxhZU8	Ellie Goulding	Love Me Like You Do
Gustavo Trebien	QeXY678B7pM	Joan Osborne	One of Us
Gustavo Trebien	SrnCvD2OHIQ	Ronan Keating	When You Say Nothing at All
Hannah Trigwell	MGs2f1ncMgA	Sam Smith	Stay With Me
Harry Pane	MY7Qgth1ieE	Kygo	Firestone
Henry Capelossi	y6qVcW8BM80	Whitesnake	Is This Love
Iman	BrcPR1PpNkM	Rihanna	Consideration
isabelle alexandra	89mfGEcrsWc	50 Cent	21 Questions
Jackie D Williams	Rew1txwh1HI	Disclosure	Latch
Jacob Banks	pv4ngbYebhg	Marvin Gaye	Let’s Get It On
Jake Isaac	cb9FO1koPyo	Bon Iver	I Can’t Make You Love Me
Jamé Forbes	IJFu-S47IoY	Gorillaz	Feel good inc.
James Craise	Ic-nF91qjJ8	Jessie Ware	Night Light
James Robb	t7Y3T23or7g	Pixie Lott	Nasty
Jamie Grey	_GgowRVILcU	Macklemore	Can’t Hold Us
Jamie Grey	oc8Efc5e58U	Raleigh Ritchie	Stronger Than Ever
Janick Thibault	cdKaO1qp4jk	A Day To Remember	All I Want
Janick Thibault	Hafa88i8E5Y	Blink-182	Bored To Death
Janick Thibault	IKWD_wG0S3s	Bring Me The Horizon	Drown
Janick Thibault	pNqpwTh-YE8	Journey	Don’t Stop Believing
Janick Thibault	7p9dKJXmS8Y	Linkin Park	Final Masquerade

Janick Thibault	_nTcTMY7MFg	Rihanna	We Found Love
Jasper Storey	tAIz2mo0VnM	Sam Smith	Writing's On The Wall
JaysonGW	hECTRAAgMRU	Bee Gees	How Deep Is Your Love
JC Villafan	qTDIvErbc8Q	Ne-Yo	Let Me Love You
Jenn Fiorentino	9g2JRk0OCkQ	The Offspring	Want You Bad
Jennifer Hannah	eLLvkJxMcFE	Eric Clapton	Tears In Heaven
Jeremy Ornelas	oqZTOM8HKu4	Temple of the Dog	Hunger Strike
Jeremy Passion	6rJnZRnRE4I	Gotye	Somebody That I Used To Know
Jess Greenberg	0gt5cEY6PKw	Extreme	More than words
Jess Greenberg	F6DusvaRmyY	Led Zeppelin	Stairway To Heaven
Jess Greenberg	fk8oBPatupQ	Oasis	Wonderwall
Jess Greenberg	0CZuZZ17mck	RHCP	Under the bridge
Jess Thristan	SRz85AArIoI	Taylor Swift	Blank Space
Joe Muscatello	aSLshMDX1EY	Prince	Purple Rain
John Mayer	chLFi2cFxzo	John Mayer	Queen Of California
Jon Prucha	eCq8vju2ndM	Lady Gaga	Paparazzi
JP Cooper	9EdLOMSBFDU	Rihanna	Stay
Katie Nicholas	IvWnNF0xu5M	Kacey Musgraves	Follow Your Arrow
Kina Grannis	uHi9qvaIUoI	Adele	Rolling In The Deep
Kina Grannis	qb4LjFtathI	Coldplay	Magic
KjGydell	tYHGvgdET7Y	Jake Bugg	Seen It All
Kristie Killick	Y-H6I2_OIEM	Justin Bieber	Love Yourself
KYKO	VW1ZDdseC_Q	Lionel Richie	All Night Long
Kyra	7psy1qoiFLU	Jessie J	Bang Bang
Kyra	ie06vnd9488	Miguel	Do You
Laura Zocca	gLbk4OnDYjM	Ellie Goulding	Beating Heart
Lauren Thalia	tsna7C-LRPg	Zara L.&MNEK	Never Forget You
LaurenBonnell24	SFJqpgHOnvI	Justin Bieber	Sorry
Letisha Gordon	6JSkkdXNZcw	K. Michelle	Love 'Em All
Lianne Kaye	P1BoIGK7uDc	Imagine Dragons	Demons
Lianne Kaye	QSjfPWfPmgo	John Newman	Blame
Lilly Ahlberg	g-FdYrstOWo	Tove Lo	Habits (Stay High)
Linah Is	jmi0SAscvw0	Major Lazer	Lean On
Lindsey Saunders	tUiDg26s0o0	The Beatles	Something
Lisa Wright	8hFXEy603Iw	The Neighbourhood	RIP 2 My Youth
Lovelle	2d7o14FqaKw	Katy Perry	Legendary Lovers
Lucy Kilner	J5YZRYccsPk	OneRepublic	Counting Stars
Luke Duffett	yqgj7v_nDnc	The Cure	Friday I'm in Love

Luke Towler	jB_a7DtqvMg	5 SoS	She Looks So Perfect
Mackenzie J.	Rn00vAIcnR4	Adele	Hello
Mackenzie J.	3IvM02BSMxI	Ed Sheeran	All Of The Stars
Mackenzie J.	LLAhxpVCCe0	Ed Sheeran	Don’t
Mackenzie J.	53TI_E6WcL8	Tracy Chapman	Fast Car
Magdalena Wolk	OspgP3hpcxU	Banks	Bette
Mark Asari	gQDSWoxnvFU	Rihanna	FourFive Seconds
Matt Wills	9jVif6czkAc	AlunaGeorge	You Know You Like It
Mel Plant Music	a8-t1FVxIQQ	Bob Dylan	Hurricane
Mike Massé	jHzmi487z8U	David Bowie	Heroes
Mike Massé	ixOj_83IkEQ	Five for Fighting	Superman
Mike Massé	E0ELg_rjIIQ	Who	Behind Blue Eyes
Mike Peralta	nK3cRrVOxgM	5 SoS	She’s Kinda Hot
Mike Peralta	C8C5dqXX6hA	Ed Sheeran	Kiss Me
Mike Peralta	uDjQtjIxmkI	Maroon 5	Sugar
Mike Peralta	o5vAI-qb54M	The Ramones	Pet Sematary
Mila Falls	_9dXWgOrtMg	Grimes	Flesh Without Blood
Miss Sha	4NwDRn8WfMo	Rita Ora	I Will Never Let You Down
moderndaywarrior	-EEanIx2W5g	The Distillers	The Hunger
Mullally	6BKBZCRIpC4	Zayn	Pillow Talk
murasakimochi	bG4fBOiG7Lw	Nobuo Uematsu	Eyes on Me
Noah Cover	I7XEggwOabs	The Zutons	Valerie
Obadiah Parker	8ejeEBIDESc	Obadiah Parker	Hey Ya
Olivia Leisk	ZiLIAYuGpGY	Taylor Swift	I Knew You Were Trouble
ortoPilot	rkz5TegISQA	Carole King	You’ve Got A Friend
Patrick Carroll	oDkIp4pveXY	Survivor	Eye of the tiger
Phebe Edwards	K0If3_Dw3vI	Adele	All I Ask
Philippa June	Vgw_bwns_vU	Beyonce	Runnin’
Phoebe Katis	IzWIBVH72XY	OMI	Cheerleader
Project Alfie	RfLDRHOBzNE	James Bay	Let It Go
Ray Lamontagne	-YAd9hN4eew	Bee Gees	To Love Somebody
Rebecca Shearing	wd17jJQwIiU	Beyonce	Resentment
Rebecca Shearing	q23Kw1YTaUE	Bruno Mars	Locked Out Of Heaven
Rebecca Shearing	C5YIcZ3TaUQ	Robin Thicke	Blurred Lines
Richard De Soussa	2q2X9P8Ej1A	Rihanna	Man Down
Rick Farmer	wCSPzu55KFU	Faith No More	Easy
Rothwell	AQuIbWSqw28	Lukas Graham	7 Years
Rufio Summers	eEaCcUobUHM	Adele	Skyfall

Rukhsana Merrise	ip80MBQR5o8	Grades	Crocodile Tears
Sam Brett	D4vp0T5TNIU	Charli XCX	Doing It
Sasha Keable	jd1lO4sy6uM	Jhené Aiko	The Worst
Scarlet Baxter	dy1UOLBo4oo	Years&Years	King
Shannon Saunders	B09_gP0-a3I	Beyonce	Mine
Shannon Saunders	SJEiDMFPP70	Frank Ocean	Swim Good
Shaun Colwill	bJ6zMrHWLTU	Marlon Roudette	When The Beat Drops Out
Sofia	rq-RbqSXpp8	Adele	Send My Love (To Your New Lover)
Sofia	A54wJ0zxlal	AMY WINEHOUSE	Back to black
Sophia Alexa	Lfdbv57x4gI	OutKast	Hey Ya!
Sophia Alexa	kiuSmD-abxo	Sam Smith	La La La
Stephanie Rainey	Y5WoWbPI7fs	Beyonce	Runnin'
Storm	BudstJ0x-MI	Years&Years	Shine
Sylvia Mwenze	FINdwdYiljM	Jazmine Sullivan	Stupid Girl
Tom Bem	AlzPxgRW9hQ	Justin Bieber	Sorry
Tori Kelly	7_3hKVxOcRI	Justin Timberlake	Suit&Tie
Travis Cormier	G02QOSaxH-w	Bon Jovi	Always
Tristan Prettyman	AB2ukG9t7cg	Tristan Prettyman	I Was Gonna Marry You
Val Maugenest	nlSVPZ8SdB0	PARTYNEXTDOOR	Come&See Me
VintageJamesT	ykPc_6BVam4	Guns N' Roses	Don't Cry
Violet Skies	7fM1sxElshY	Ed Sheeran	Thinking Out Loud
Will Gittens	W_ydnsr0bm4	Justin Bieber	Love Yourself
Yo Preston	MSmy9zl_12g	Jason Derulo	Trumpets
Zach London	SzUEf_EpFIs	J. Cole	Workout
Zack Knight	nu_eVkPg9s0	OneRepublic	If I Lose Myself
Zella Day	PndWEWfaCPs	Zella Day	Seven Nation Army

Table A.2: List of piano songs

Cover Artist	**YouTube Link**	**Original Artist**	**Song Name**
Abi Alton	J2ZAo05TS9s	Jamie Lawson	Wasn'T Expecting That
Abi Alton	tUQnG3SiVtk	Onerepublic	I Lived
Ben Schuller	ajmtkfbBKac	Rihanna	Work
Benedict	VzfwENTrev4	Disclosure	White Noise
Beth	4zu9kAjXHjY	Alan Walker	Faded
Beth	EsDfwqzOJRc	Calvin Harris	How Deep Is Your Love
Beth	GoSW-I0FaXc	Justin Bieber	What Do You Mean?
Boyce Avenue	8SF1Wt__W6g	Ellie Goulding	Love Me Like You Do
Boyce Avenue	LWgqWuJG5Jg	Adele	Hello

Carrie Haber	Sivw6a4KiW4	Calvin Harris	I Need Your Love
Ellysa Rose	U4_IPPMXbYg	Justin Bieber	Sorry
Emi Mcdade	ndgC-ZquPRM	Kodaline	All I Want
Erika David	9HPViNkJNMc	The Wanted	Glad You Came
Hope	Wdd764CDmh0	Birdy	Wings
Jackie D Williams	svUupqIauiY	Passenger	Let Her Go
Jacob Wellfair	YAo-CSE-fbE	Birdy	Wings
James Robb	OS-YLg4CFXQ	Nick Jonas	Jealous
Jamie	3Ie3V5TiQgI	Jamie Xx	Loud Places
Jody Brock	p1LqIxPWeBw	The Lumineers	Stubborn Love
Kelli-Leigh	XV6xI6w5HWo	Jessie Ware	Wildest Moments
Lauren Aquilina	yFtZMmETxuU	Coldplay	Magic
Lauren Faith	XsCHyk_PE5U	Pharrell Williams	Lost Queen
Mallie	XMXXsG0wHIo	Ed Sheeran	Give Me Love
Martin Luke Brown	ThqLMTwJ9iE	Emeli Sandé	Free
Mexirass	Fp4vxDMIJWM	Tom Petty	Free Fallin'
Mike Massé	-YAu-sGIpwQ	The Beatles	Let It Be
Ollie Sloan	08JOHQ8hgIM	Ellie Goulding	How Long Will I Love You
Pharella	ZRMNzgitpIY	Selena Gomez	The Heart Wants What It Wants
Quigley	LBeeBxwaAOM	Onerepublic	Counting Stars
Raphaella	iGV0uwP_10w	Imagine Dragons	Radioactive
Raphaella	svSuQGDyed8	Justin Bieber	As Long As You Love Me
Rebecca James	8LZTcgIIJbk	Sam Smith	Writing'S On The Wall
Rebecca Shearing	Fr4xgCUdHfk	Adele	All I Ask
Rebecca Shearing	JEctXK2YPFA	Flume	Never Be Like You
Sally Caitlin	3L0SIkNS4Ig	Justin Bieber	Sorry
Sarah Close	Fq4huCq56RM	Owl City	Good Time
Sid Batham	Y0ukDuQDGV4	Lianne La Havas	Gone
Sofia	m-Ia4Km8Kqc	John Legend	All Of Me
Tom Prior	8R5KKTvjBzo	Naughty Boy	Home

Appendix B

Language Model Source

Database Artist List:

- 50 Cent
- 5 Seconds Of Summer
- A Day To Remember
- Adele
- A Great Big World
- Alan Walker
- Alessia Cara
- Al Green
- Alunageorge
- Amy Winehouse
- Bad Religion
- Banks
- Bee Gees
- Beyonce
- Birdy
- Blink-182
- Bob Dylan
- Bon Iver
- Bon Jovi
- Bouncing Souls
- Bring Me The Horizon
- Bruno Mars
- Calvin Harris
- Carole King
- Charli XCX
- Chris Stapleton
- Cody Simpson
- Coldplay
- Creed
- David Bowie
- Disclosure
- Ed Sheeran
- Ellie Goulding
- Elton John
- Emeli Sandé
- Eric Clapton
- Extreme
- Faith No More
- Five For Fighting
- Florida Georgia Line
- Flume
- Foo Fighters
- Frank Ocean
- Gnarls Barkley
- Gorillaz
- Gotye
- Grades
- Grimes
- Guns N' Roses
- Imagine Dragons
- Jake Bugg
- James Bay
- Jamie Lawson
- Jamie XX
- Janet Jackson
- Jason Derulo
- Jazmine Sullivan
- J. Cole
- Jennifer Hudson
- Jessie J
- Jessie Ware
- Jhené Aiko
- Joan Osborne
- John Legend
- John Lennon
- John Mayer
- John Newman
- Journey
- Justin Bieber
- Justin Timberlake
- Kacey Musgraves
- Katy B
- Katy Perry
- Kiesza
- K. Michelle
- Kodaline
- Kwabs
- Kygo
- Lady Gaga
- Led Zeppelin
- Leonard Cohen
- Lianne La Havas
- Linkin Park
- Lionel Richie
- Lorde
- Lukas Graham
- Macklemore
- Magic!
- Major Lazer
- Marlon Roudette
- Maroon 5
- Marvin Gaye
- Michael Jackson
- Miguel
- Naughty Boy

- Ne-Yo
- Nick Jonas
- Nirvana
- Nobuo Uematsu
- Oasis
- Obadiah Parker
- Omi
- Onerepublic
- Outkast
- Owl City
- Partynextdoor
- Passenger
- Pearl Jam
- Peter Cetera
- Pharrell Williams
- Pink Floyd
- Pixie Lott
- Pixies
- Prince
- Queen
- Radiohead
- Raleigh Ritchie
- Red Hot Chili Peppers
- Rihanna
- Rita Ora
- Robin Thicke
- Rolling Stones
- Ronan Keating
- Sam Smith
- Scorpions
- Seal
- Selena Gomez
- Shawn Mendes
- Sia
- Stevie Wonder
- Survivor
- Taylor Swift
- Temple Of The Dog
- The Beatles
- The Cure
- The Distillers
- The Doors
- The Lumineers
- The Neighbourhood
- The Offspring
- The Police
- The Ramones
- The Wanted
- The Zutons
- Tom Petty
- Tonic
- Tori Amos
- Tove Lo
- Tracy Chapman
- Tristan Prettyman
- U2
- Whitesnake
- Who
- Years & Years
- Zara Larsson & MNEK
- Zayn
- Zella Day

Complementary List:

- ACDC
- Adam Lambert
- Artic Monkeys
- Avril Lavigne
- Blur
- Celine Dion
- Demi Lovato
- Elvis Presley
- George Michael
- Green Day
- Jennifer Lopez
- Joe Cocker
- Johnny Cash
- Johnny Marr
- Kylie Minogue
- Leona Lewis
- Lily Allen
- Madonna
- Melanie C
- Michael Buble
- Miley Cyrus
- Muse
- Neil Diamond
- Noel Gallagher
- Ozzy Osbourne
- Paul McCartney
- Phil Collins
- Pitbull
- Robbie Williams
- Rod Stewart
- Shania Twain
- Stereophonics
- Sting
- Take That
- The Black Eyed Peas
- The Stone Roses
- Victoria Beckham
- Wings
- Wolf Alice

Appendix C

Implementation of the project

Exist several technical decisions that should be taken when an ASR system is designed and all of these decision may affect directly on how it will be implemented. For this project, the language code selected for implementing each stage of the project was **Python 2.7** because the flexibility and simplicity it offers. Moreover, **Kaldi** [36] was the chosen ASR toolkit selected over HTK [4] toolkit because, Kaldi successfully integrate several DNN models features needed for this project.

The construction and implementation of this project was separated in several stages that comprise an specific task in the research. Following, the implementation of this stages will be explained.

C.1 ACOMUS1 Implementation

The ACOMUS construction started with the preselection of the 240 song that comprising the current corpus. In this process, several songs that have duet or polyphonic instrumentation were discarder for this first version of the corpus. With a python tool, the list of the songs was separated into three database files with an **ID** for speakers and songs. After the song list separation into id files, several sub-stages were implemented. Firstly, the python tool **download_audio_database.py** that read a batch file with the whole song links was implemented. This tool call the library **youtube-dl** and download the raw audio file of each song and save it with the corresponding name.

Secondly, the annotation implementation was separated into seven subtask where each of them perform a small step becoming the code easy to maintain.

1. Annotate using **annotate.py**:
   Interactive tool that allows to play an unsegmented audio and manually record the start and end points of each utterances. The output is a raw annotation file in a json format.

2. Produce utterance level json file **segment.py**:
   Takes the previous raw annotation file and produces an utterance level segmentation json file.

3. Turn lyrics into json **jsonify_lyric.py**:
   Take the cover lyric file and produce a json structured file. It is mandatory that each lines in the lyrics file match with the sentences in the segments file.

4. Add lyrics to main json **add_lyrics.py**:
   This tool merge the segment and the lyrics json files.

5. Refine annotations using **refine.py**:
   This tool allows to refine the start and end point of each sentences. Also, allows to change and refine the exact word that the singer said in the sentence.

6. Extract utterance segments **extract.py**:
   Using the refined json lyrics from above, takes the raw audio file and extract it into several isolated utterance wav files.

7. Generate the text file of song **utterance_sentences.py**: Using the same refined json lyrics, produce the song version of the {TEXT} file necessary for Kaldi toolkit.

The update state and the details of the construction of the ACOMUS1 corpus are saved in the repository `http://github.com/groadabike/ACOMUS`.

C.2 Language Model & Acoustic Model Implementing

For the implementation of the LM and construction of the 5K words vocabulary, four python tools were coded.

- Download Artist Discography **down_lyrics.py**:
  This tool download from **lyrics.wikia.com** the lyrics discography from a list of artist names. For this project, the list used was the list displayed in Appendix B.

- Collate each track into one file **join_lyrics.py**:
  Collate all the lyrics downloaded from previous tool into one *collection* file.

- Lexify the collection lyrics **lexify_lyrics.py**:
  Analyse the content of the lyrics collection file and perform the normalization process.

- Create 5K vocabulary **create_vocabulary.py**:
  Using the CMUdict lexicon file this tool generate the 5K words vocabulary in a **lexicon.txt** file needed by kaldi.

C.3 Cooking the Recipe

For the final kaldi recipe implementation, this project follows the structure of the fifth version of the toolkit. For the preparation of the mandatory files two python tools were created.

- **create_corpus.py**:
  This tool takes the final lexify collection file from the LM process and construct the kaldi's LM corpus file.

- **data_info_files.py**:
  Tool that generate the Kaldi's files:

spk2gender	<speaker ID> <gender>
wav.scp	<utteranceID> <full_path_to_audio_file>

text	<utteranceID> <text_transcription>
utt2spk	<utteranceID> <speakerID>

The recipe used and its specifications are available in the repository `http://github.com/groadabike/ASR-music`.

Appendix D

Experiments Results

Table D.1: Detail of the insertions, deletions and substitutions words decoding the baseband experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1	3689	Mono	67	1422	1929
		Tri1	88	1336	1960
		Tri2B	84	1407	1898
		Tri3B	59	1474	1779
		DNN	113	1196	2018
Run2	4530	Mono	106	1792	2373
		Tri1	87	2035	2148
		Tri2B	87	2123	2064
		Tri3B	91	1983	2185
Run3	4378	Mono	120	1526	2481
		Tri1	97	1954	2059
		Tri2B	74	1985	2064
		Tri3B	77	1883	2109
Run4	4512	Mono	136	1572	2519
		Tri1	101	1890	2273
		Tri2B	108	1884	2275
		Tri3B	88	1929	2202

Table D.2: Detail of the insertions, deletions and substitutions words decoding the pitch augmentation experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1.1	3689	Mono	83	1356	1975
		Tri1	97	1269	1980
		Tri2B	74	1322	1951
		Tri3B	66	1290	1898
		DNN	166	975	2128
Run2.1	4530	Mono	83	1799	2380
		Tri1	108	1828	2289
		Tri2B	109	1897	2251
		Tri3B	164	1612	2386
Run3.1	4378	Mono	99	1631	2400
		Tri1	80	1985	2054
		Tri2B	105	1802	2203
		Tri3B	152	1496	2366
Run4.1	4512	Mono	89	1759	2390
		Tri1	103	1798	2339
		Tri2B	98	1853	2299
		Tri3B	177	1485	2501

Table D.3: Detail of the insertions, deletions and substitutions words decoding the tempo augmentation experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1.2	3689	Mono	59	1474	1908
		Tri1	77	1357	1996
		Tri2B	78	1406	1917
		Tri3B	98	1277	1953
Run2.2	4530	Mono	70	1851	2294
		Tri1	94	1997	2229
		Tri2B	121	1873	2319
		Tri3B	152	1710	2389
Run3.2	4378	Mono	75	1934	2127
		Tri1	123	1764	2229
		Tri2B	70	2047	2016
		Tri3B	128	1752	2184
Run4.2	4512	Mono	108	1657	2475
		Tri1	93	1920	2268
		Tri2B	121	1770	2384
		Tri3B	121	1776	2321

Table D.4: Detail of the insertions, deletions and substitutions words decoding the pitch + tempo without combination augmentation experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1.3	3689	Mono	91	1350	1924
		Tri1	112	1114	2134
		Tri2B	52	1538	1765
		Tri3B	104	1192	1972
Run2.3	4530	Mono	66	1822	2347
		Tri1	72	1919	2267
		Tri2B	128	1663	2459
		Tri3B	128	1719	2365
Run3.3	4378	Mono	83	1834	2223
		Tri1	74	1931	2121
		Tri2B	94	1906	2129
		Tri3B	95	1765	2181
Run4.3	4512	Mono	120	1675	2405
		Tri1	101	1797	2335
		Tri2B	117	1790	2355
		Tri3B	125	1642	2406

Table D.5: Detail of the insertions, deletions and substitutions words decoding the pitch + tempo with combination augmentation experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1.4	3689	Mono	68	1522	1834
		Tri1	90	1438	1837
		Tri2B	95	1383	1902
		Tri3B	90	1290	1887
		DNN	191	1010	2173
Run2.4	4530	Mono	84	1707	2478
		Tri1	188	1467	2605
		Tri2B	108	1848	2279
		Tri3B	144	1616	2419
Run3.4	4378	Mono	53	1891	2215
		Tri1	73	2033	2034
		Tri2B	79	1927	2154
		Tri3B	193	1306	2567
Run4.4	4512	Mono	109	1775	2364
		Tri1	128	1714	2373
		Tri2B	152	1687	2386
		Tri3B	134	1594	2436

Table D.6: Detail of the insertions, deletions and substitutions words decoding the pitch as a new same gender singer augmentation experiments with development dataset.

Experiment	Total Words	Model	Insertion	Deletion	Substitution
Run1.1S	3689	Mono	42	1674	1708
		Tri1	54	1589	1716
		Tri2B	61	1464	1823
		Tri3B	60	1439	1791
Run2.1S	4530	Mono	87	1768	2386
		Tri1	95	1893	2279
		Tri2B	88	1899	2230
		Tri3B	131	1776	2304
Run3.1S	4378	Mono	94	1601	2420
		Tri1	70	1918	2154
		Tri2B	97	1754	2267
		Tri3B	131	1712	2233
Run4.1S	4512	Mono	64	1881	2252
		Tri1	103	1820	2295
		Tri2B	122	1830	2322
		Tri3B	168	1600	2395

Table D.7: Detail of the insertions, deletions and substitutions words decoding the guitar test set.

Model	Total Words	Insertion	Deletion	Substitution
Tri3B	9206	200	3283	4876
DNN sMBR		314	2873	5087

Table D.8: Detail of the insertions, deletions and substitutions words decoding the piano test set.

Model	Total Words	Insertion	Deletion	Substitution
Tri3B	9268	333	2516	5185
DNN sMBR		429	2227	5184