Learning Idioms and Slang using a Deep Learning approach with Speech Recognition

Davide Berdin
{davide.berdin.0110}@student.uu.se

Master Thesis
Department of Information Technology

November 7, 2015

Abstract

To my family that never stopped believing in me

Acknowledgements

Contents

List of Figures

Chapter 1

Introduction

The pronunciation is one of the hardest part of learning a language among all the other components, such as grammar rules and vocabulary. To achieve a good level of pronunciation, non native speakers have to study and constantly practice the target language for a incredible amount of hours. In most cases, when students are learning a new language, the teacher is not a native speaker in which implies that the pronunciation may be influenced by the country where he or she comes from, since it is a normal consequence of second learning language [11]. In fact, [12] states that the advantages of having a native speaker as a teacher, lies in the superior linguistic competences, especially the usage of the language more spontaneously in different communication situations. Pronunciation falls into those competences underlying a base problem in teaching pronunciation at school.

The basic questions we tried to answer in this work are:
1) Why is pronunciation so important?
2) What are the most effective methods for improving the pronunciation?
3) What is the research state-of-art and what can we do to make it better?

The first question is fairly easy to answer. There are two reasons to claim why pronunciation is important: *(i)* it helps to acquire the target language faster and *(ii)* being understood. Regarding the first point, earlier a learner masters the basics of pronunciation, the faster it will become fluent. The reason is because *critical listening* with a particular focus on hearing the sounds will lead to gain fluency in speaking the language. The second point is **crucial** when working with other people, especially at these days where both in school and business the environment is often multicultural. Pronunciation mistakes may lead the person to being misunderstood affecting the results of a project for example.

With these statements in mind, [13] gives suggestions on how a learner can effectively improve the pronunciation. Four important ways are depicted: *Conversation* is the most relevant approach to improve pronunciation, although, a supervision of an *expert guidance* that corrects the mistakes is fundamental during the process of learning. At the same time, learners have to be pro-active to have conversation with other native speakers in such a way to constantly practicing. *Repetitions* of pronunciation exercises is another important factor that will help the learner to be better in speaking. As last, *Critical listening*, that we also mentioned above, amplify the opportunity of learning the way native speakers pronounce words. In particular, for a learner is important to understand the difference when he or she is pronouncing a certain sentence with the one said by the native speaker. This method is very effective and is important for understanding the different sounds of the language and how a native speaker is able to reproduce them [14].
An important factor while learning a second language is to have a feedback about improvements. Teachers are usually responsible to judge the way the learners' progress. In fact, when teaching pronunciation, often it is used to draw the intonation and the stress of the words in such a way that the learner is able to see how the utterances should be pronounced. The *British Council* shows this practice [15]. The usage of visual feedbacks is the key of learning pronunciation and it is the main feature of this research.

In the computer science field, some works have been previously done regarding pronunciation. For instance, [16] helps learners to acquire the tonal sound system of Mandarin Chinese through a mobile game. Another example is [17] in which the application provides a platform where learners of Chinese language can interact with native speakers and challenging them to a competition of pronunciations of chinese tones.

The idea behind this project is based on the fact that people need to keep practicing their pronunciation to have a significant improvement as well as they need immediate feedbacks to understand if they are going in the right direction or not. The approach we used is based on these two factors and we designed the system to be as useful and portable as possible. The mobile application is where the user will test the pronunciation, a server through a machine learning technique will compute the similarity between the user's pronunciation and the native speaker's one and the results will be displayed on the phone.

We started collecting data from *American Native Speakers* in which we asked to pronounced a set of most used idioms and slangs. Each candidate had to repeat the same sentence four times trying to be as much consistent as possible. After we gathered the data, a preprocessing step is due since we are seeking for specific features such as voice-stress, accent, intonation and formants. This part is actually divided in two parts as well, where we used **PRAAT**[18] to extract the features from each audio file and then we used **MATLAB**[1] to average the four audio signals that each candidate said. After the averaging step, there is a smoothing process in which we applied the *Exponential Filter* to each signal. A *resampling* was necessary since not all the signals have the same length in time. The next step consisted in the so called **force alignment** where for each audio file we extract the *phonemes* that have been pronounced.
The same treatment is given to the audio signals given by the final user in order to compare the pronunciation.

The machine learning part is composed by a *deep neural network* trained using a statistical modeling method called **conditional random fields**. The system is trained using the features extracted during the data processing phase, coupled with the phonemes retrieved using the force alignment. The phonemes are used as labels for the classification in which we indeed predict. To estimate the overall error between the native pronunciation and the user, we use a method called **Word Error Rate** (WER), a common method metric for measuring the performance of a speech recognition system. In addition to *WER*, we used **Dynamic Time Warping** (DTW) to evaluate the similarity between features. These results will allow the user to have a better understanding of the differences for each characteristic.
DTW is not a part of the neural network, but it's a standalone algorithm.

When the server has computed the similarity, a score with the evaluation of voice-stress, accent, intonation and formants is given to the user. Those features are not the only one available. In fact, the application is capable to show the graph for each feature of the user's pronunciation and the one of the native speaker. In this way the user has a clear understanding of how he or she should **adjust** the way the utterance should be pronounced accordingly.

To measure the accuracy of our model, we asked 10 native speakers to rate the pronunciation of 10 non native speakers on 5 different sentences. The range of evaluation was between 0 and 10. From here we averaged all the rates and compared with the results obtained by the automatic system.

[1] http://www.mathworks.com

Chapter 2

Sounds of the General American English

In the *General American English* there are 41 different sounds in which can be structured by the way they are produced. In 2.1 is shown the kind of sounds with the respective number of possible productions. Each type will be described into a dedicated section of this thesis. An important factor is the way of how the *constriction* of the flow of air is made. In fact, to distiguish between *consonants*, *semivowels* and *vowels*, the *degree* of constriction is checked. Instead, for *sonorant* consonants the air flow is continuous with no pressure. *Nasal* consonants have an occlusive consonant made with a lowered velum allowing the airflow in the nasal cavity [19]. The *continuant* consonants are produced without blocking the airflow in the oral cavity.

Type	**Number**
Vowels	18
Fricatives	8
Stops	6
Nasals	3
Semivowels	4
Affricates	2
Aspirant	1

Table 2.1: Type of English sounds

2.1 Vowel production

Generally speaking, when a vowel is pronunced, there is no air-constriction in the flow. This means that the ariculators like the tongue, lips and the uvula do not touch allowing the flow of air from the lungs. The consonants instead have another pattern when producing them. Moreover, to produce each vowel, the mouth has to make a different shape in such a way that the resonance is different. 2.1 shows the way the mouth, the jaw and the lips are combined in a such a way to produce the acoustinc sound of a vowel.

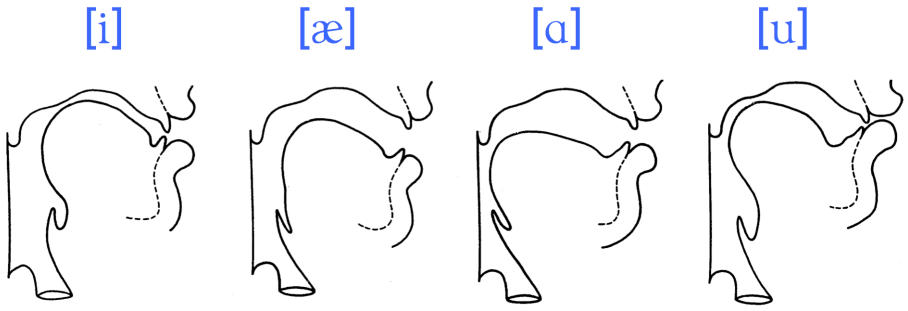

Figure 2.1: Vowels production [1]

2.1.1 Vowel of American English

There are 18 different vowels in American English that can be grouped by three different sets: the **monopthongs**, the **diphthongs**, and the **schwa's** - or reduced vowels.

/iʸ/	iy	beat	/ɔ/	ao	bought	/ɑʸ/	ay	bite
/ɪ/	ih	bit	/ʌ/	ah	but	/ɔʸ/	oy	Boyd
/eʸ/	ey	bait	/oʷ/	ow	boat	/ɑʷ/	aw	bout
/ɛ/	eh	bet	/ʊ/	uh	book	[ə]	ax	about
/æ/	ae	bat	/u/	uw	boot	[ɨ]	ix	roses
/ɑ/	aa	Bob	/ɝ/	er	Bert	[ɚ]	axr	butter

Figure 2.2: Example of words depending on the group [1]

The first column shows some examples of monopthongs. A *monopthong* is a clear vowel sound in which the utterance are fixed at both the beginning and at the end. The central part of the picture represents the dipthongs. A *dipthong* is the sound produced by two vowels when they occur within the same syllable [20]. In the last column are depictes some examples of reduced vowels. *Schwa's* refers to the vowel sound that stays in the mid-central of the word. In general, in the english language, the schwa is found in unstressed position [21].

2.1.2 Formants

A *formant* is the resonant frequency of a vocal track that resonate the loudest. In a spectrum graph, formants are represented by the peaks. In 2.3 it is possibile to see how the three first formants are defined by the peaks. The pictures is the *envelope* of a spectogram of the vowel **[i]**. Frequencies are the most relevant information to determine which vowel has been pronounced. In general, within a spectrum graph there may be a different number of formants, although the most relevant are the first three and they are named **F1**, **F2** and **F3**.

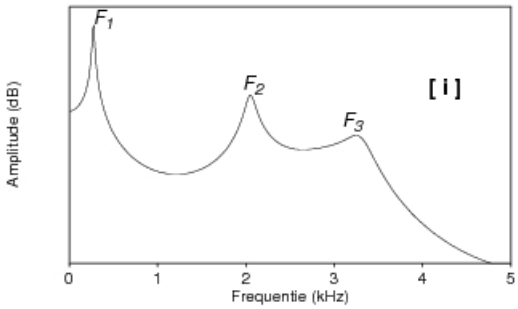

Figure 2.3: Spectral envelope of the [i] vowel pronunciation. F1, F2 and F3 are the first 3 formants [2]

The frequencies produced by the formants are highly dependent on the tongue position. In fact, formant *F1*'s frequencies are produced when the tongue is either in a *high* or *low* position, whereas formant *F2* whene the tongue is in either *front* or *back* position and formant *F3* when the tongue is doing *Retroflexion*. **Retroflextion** is more present when pronouncing the consonant *R*

2.1.3 Vowel duration

The duration of a vowel is the time that taken when pronounicing it. The duration is measured in *centiseconds* and in English[1] the different lengths are defined by certain rules. In general, the length of *lax vowels* such as /ɪ e æʌ ɒ u ɘ/ are short whereas *tense vowels* like /iː ɑː ɔː uː ɜː/ including dipthongs /eɪ aɪ ɔɪ əʊ aʊ ɪə ea ʊə/ have a variable lentgh but longer than lax vowels [3]. In 2.4 is shown an example of time-length of some vowels. In General American English, the length of vowels are not as distinctive as in the *RP*[2] pronunciation. In some American accents, to express an emphasis the length of vowels can be extended.

[1] In Icelandic as well
[2] More commonly referred as the Standard English in the UK

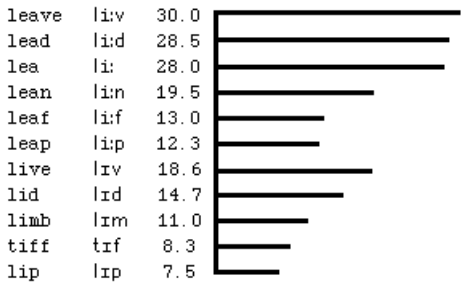

Figure 2.4: RP vowel length [3]

2.2 Fricative Production

A **fricative** is a consonant sound that is produced by narrowing the cavity causing a friction as the air goes through it [22]. There are eight fricatives in American English divided in two categories: *Unvoiced* and *Voiced*. These two categories are often called *Non-Strident* and *Strident* that means that there is a constriction behind the alveolar ridge.

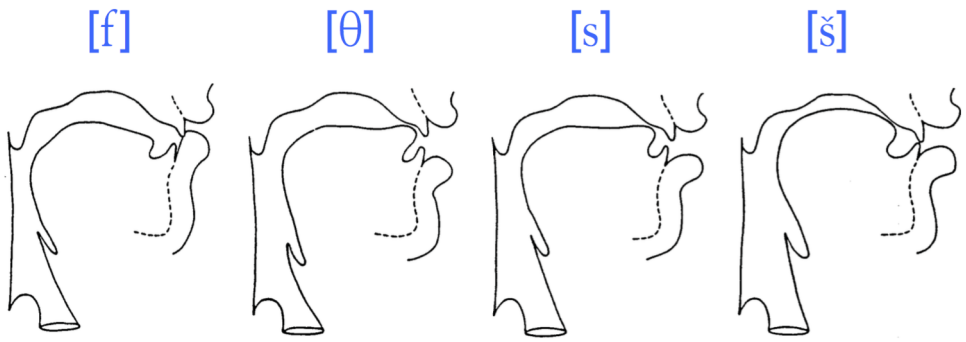

Figure 2.5: Fricative production [1]

In 2.6 it is possible to see some examples of these two categories. Each consonant also belongs to a specific articulation position. In fact, each figure in 2.5 represents a specific articulation position. From left to right we have: *Labio-Dental* (Labial), *Interdental* (Dental), *Alveolar* and *Palato-Alveolar* (Palatal).

Type	Unvoiced			Voiced		
Labial	/f/	f	fee	/v/	v	v
Dental	/θ/	th	thief	/ð/	dh	thee
Alveolar	/s/	s	see	/z/	z	z
Palatal	/š/	sh	she	/ž/	zh	Gigi

Figure 2.6: Fricative examples of productions [1]

2.3 Affricate Production

An **affricate** consonant is produced by stopping the airflow first and then release it similar to a fricative. The result is also considered a *turbolence noise* since the produced sound has a suddent release of the constriction. In English there only two affricate phonemes, as depicted in 2.7.

Voiced			Unvoiced		
/ǰ/	jh	judge	/č/	ch	church

Figure 2.7: Affricative production [1]

2.4 Aspirant Production

An **aspirant** consonant is a strong outbreak of breath produced by generating a turbolent airflow at glottis level. In American English exists only one aspirant consonant and it is the /h/, for instance in the word *hat*.

2.5 Stop Production

A **Stop** is a consonant sound in which the oral cavity is blocked in such a way that the airflow ceases. The stop consonant is also known as *plosive* which means that it is an oral *occlusive* sound [23]. The occlusion can come up in three different variance as shown in 2.8: from left to right we have a *Labial* occlusion, the *Alveolar* occlusion and the *Velar* occlusion. The pressure built up in the vocal tract, determine the produced sound depending on which occlusion is performed.

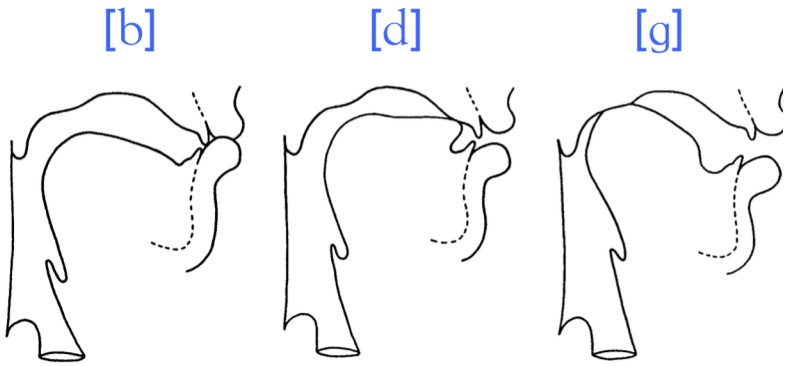

Figure 2.8: Stop production [1]

In American English there are six stop consonants, as represented in 2.9. As for the fricative consonants, the two main categories are the *Voiced* and *Unvoiced* sounds. Although, a particularity of the Unvoiced stops is that they are typically *aspirated* whereas in the Voiced ones there is a *voice-bar* during the closure movement. These two particularities are very useful where analyzing the formants because the frequencies are very well distinguished allowing a classification system to better understand the difference between stop phonemes.

Type	Voiced			Unvoiced		
Labial	/b/	b	bought	/p/	p	pot
Alveolar	/d/	d	dot	/t/	t	tot
Velar	/g/	g	got	/k/	k	cot

Figure 2.9: Stop examples of production [1]

2.6 Nasal Production

A **Nasal** is a occlusive consonant sound that is produced with a *lowered velum*, allowing the airflow to go out through the nostrils [19]. Becuase the airflow escapes through the nose, the consonants are produced with a closure in the vocal tract. 2.11 shows the three different positions to produce a nasal consonant. From left to right we have *Labial*, *Alveolar* and *Velar*.
Due to this particularity, the frequencies of nasal *murmurs* are quite similar. If we take a look on the spectrogram

in 2.10, it is possible to notice that nasal consonants have a high similarity. In a classification system, this can be a problem.

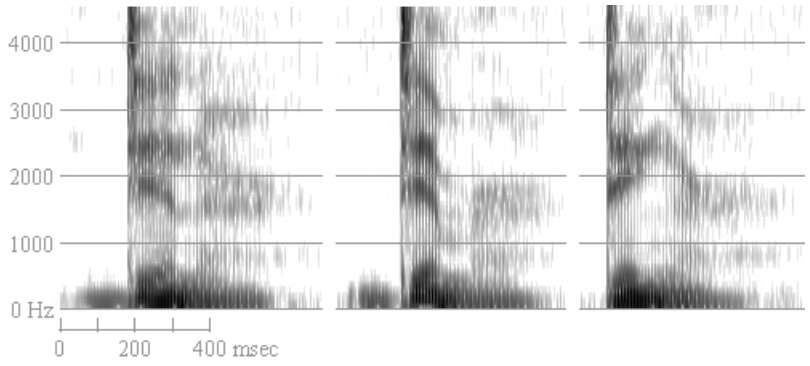

Figure 2.10: Nasal Spectrograms of "dinner", "dimmer", "dinger" [4]

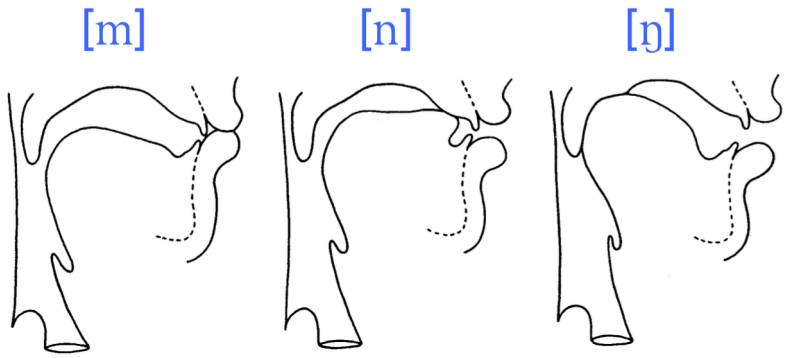

Figure 2.11: Nasal production [1]

Since the sound produced by a nasal is produced with an occlusive vocal tract, each consonant is **always attached** to a vowel and it can can form an entire syllable. Although, in English, the consonant /**ŋ**/ always occur immediately after a vowel. In 2.12 are shown some examples of nasal consonants divided by articulation position.

Type	Nasal		
Labial	/m/	m	me
Alveolar	/n/	n	knee
Velar	/ŋ/	ng	sing

Figure 2.12: Nasal examples of production [1]

2.7 Semivowels Production

A **semivowel** is a sound that is very close to a vowel sound but it works more likely as a syllable boundary rather than a core of a syllable [24]. A typical example of semivowels in English are the **y** and **w** in words *yes* and *west*. In the *IPA* alphabet they are written /j/ and /w/ and they correspond to the vowels /iː/ and /uː/ in the words *seen* and *moon*. In 2.14 there are some examples of semivowels production.
The sound is produced by making a constriction in the oral cavity without having any sort of air turbolance. To achieve that, the articulation motion is slower than other consonants becuase the laterals[3] form a complete closer combined with a tongue tip. In this way the airflow has to pour out using the sides of the constriction.

[3]They are a pair of upper teeth that are located laterally from the central incisors [25]

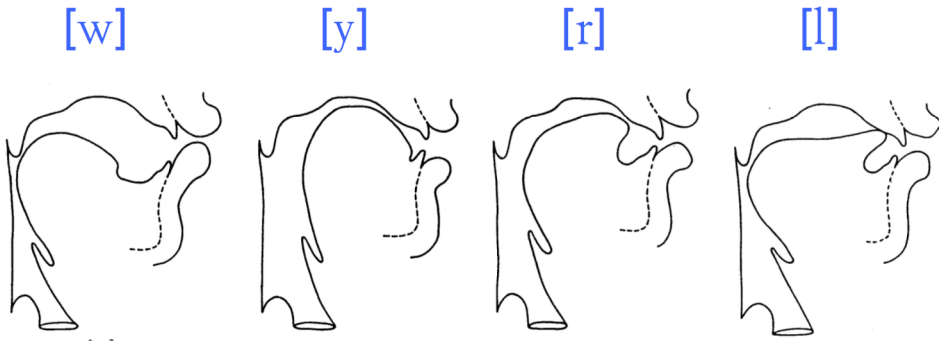

Figure 2.13: Semivowel production [1]

In American English there are four semivowels and they are depiscted in 2.13. An important fact of semivowels is that they are always close to a vowel. Although, the /l/ can form an entire syllable by itself when there is no stress in a word.

Type	Semivowel			Nearest Vowel
Glides	/w/	w	wet	/u/
	/y/	y	yet	/i/
Liquids	/r/	r	red	/ɝ/
	/l/	l	let	/o/

Figure 2.14: Semivowel examples of production [1]

Acousitc Properties of Semivowels

Semivowels have some properties that are taken into account when doing any sort of analysis. In fact, /w/ and /l/ are the semiwovels that are more confusable because both are characterized by a *low* range of frequencies for both formants *F1* and *F2*. Although, the /w/ can be distinguished by the *rapid falloff* in the F2 spectrogram whereas /l/ has more often a *high frequency energy* compared to /w/. The **energy** is the relationship between the *wavelength* and the *frequency*. So, having a high energy means that there is a high frequency value and a small wavelength [26]. The semivowel /y/ is characterized by having a very low frequency value in formant F1 and a very high in formant F2. The /r/ instead is presented with a very low frequency value of formant F3.

2.8 The Syllable

The definition of the **syllable** can be divided in two sub-definition: one from the phonetic point of view and one from the phonological point of view.
In phonetic analysis, the syllable is a basic unit of speech in which they *"are usually described as consisting of a centre which has little or no obstruction to airflow and which sounds comparatively loud; before and after that centre (...) there will be greater obstruction to airflow and/or less loud sound"* [27]. Taking the word *cat* (/kæt/) as example, the **centre** is defined by the vowel /æ/ in which takes place only a little obstruction. The surronding *plosive* consonants (/k/ and /t/) the airflow is completely blocked [28].
A phonological definition of the syllable establishes that it is *"a complex unit made up of nuclear and marginal elements"*[29]. In this context, the vowels are considered the **Nuclear** elements or syllabic segments whereas the **Marginal** ones are the consonants or non-syllabic segments [28]. Considering the word *paint* (/peɪnt/) as example, the nuclear element is defined by the diphtong /eɪ/ whereas /p/ and /nt/ are the marginal elements.

2.8.1 Syllable Structure

In the phonological theory, the syllable can be decomposed in a hierarchical structure insted of a linear one. The structure starts with the σ letter in which represents not only the root, but the syllable itself. Immediately after,

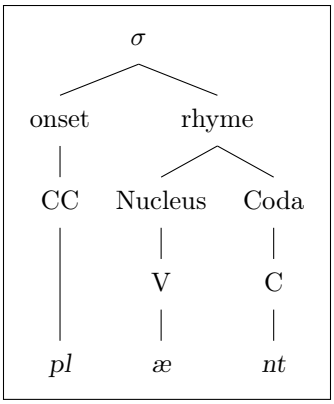

Figure 2.15: Tree structure of the word **plant**[4]

there are two *branches* called **constituents** that they represents the *Onset* and the *Ryhme*. The left branch includes any consonants that precede the vowel (or Nuclear element), whereas the right branch includes both the nuclear element and any consonants (or Marginal elements) that potentially could follow it.
Usually, the ryhme branch is further split in two other branches represented by the **Nucleus** and the **Coda**. The first on represent the nuclear element in the syllable. The second one instead, subsumes all the consonants that follow the Nucleus in the syllable [28]. In 2.15 there is a representation of the syllable structure based on the word *plant*.

[4] **C** means *Consonant* whereas **V** means *Vowel*

Chapter 3

Acoustics and Digital Signal Processing

In the past decade, digital computers have significantly helped *signal processing* to quantify a finite number of bits. The flexibility inherited from digital elements allowed the usage of a vast number of techniques in which had been not possible to implement in the past. Nowadays, digital signal processor have been used to perform multiple operations, such as *filtering*, *spectrum estimation* and many others algorithms [30].

3.1 Speech signals

The **speech** is the human way of communication. The protocol used in communication is based on a syntactic combination of different words taken from a very large vocabulary. Each word in the vocabulary is composed by a small set of vowels and consonants that combined with a phonetic units form a spoken word.
When a word is pronounced[1], a sounds is produced causing the air particles to be excited at a certain vibration rate. The source of our voice is due to the vibration of the vocal cords. The resultant signal is a *non-stationary* but it can be divided in segments since each phoneme has a common acoustic properties. In 3.1 is possible to notice how the pronounced words have a different shape as well as when the intensity of the voice is higher/lower during the pronunciation.

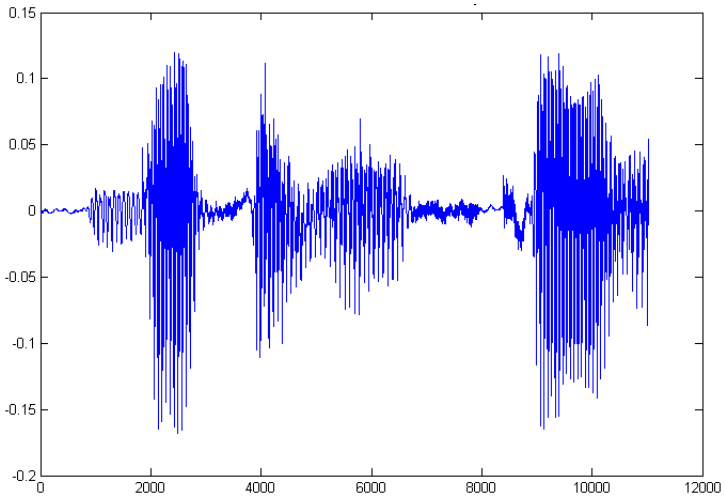

Figure 3.1: Example of a speech sound. In this case, the sentence "This is a story" has been pronounced [5]

The simplest form of sound is the *sinusoid* and it is the easiest waveform to describe because it corresponds to a **pure tone**. A pure tone consist in a waveform that consists only on one frequency. Other examples are the *cosine* or *sine* waves.

[1] 2 explains in details how phonemes are pronounced

3.1.1 Properties of Sinusoids

A sinusoid, is a simple waveform represented by a up and down movement. There are three important measures that has to be taken into consideration when defining the shape of the sinusoid: *amplitude*, *frequency* and *phase*.

Amplitude

The amplitude, from a sound point of view, corresponds to the *loudness* whereas in the soundwave it corresponds to the amount of **energy**. In general, to measure the amplitude, we use the unit called **deciBels** (dB) in which it is measured using a logarithmic scale relative to a standard sound [31].

Frequency

Frequency is the number of cycles per unit of time[2].To define cycle, we can think of an oscillation that starts from the middle line, goes to the maximum point, down to the minimum and get back to the middle point. The unit of measure of the frequency is calculated in **Hertz** (Hz). Also, if we calculate the time taken for one cycle, we estimate the so called **period**.
Frequency plays a fundamental role with the *pitch*. In fact, changing the number of oscillations but keeping the same waveform, we are able to increase or decrease the level of the pitch.

Phase

The **phase** measures the starting point position of the waveform. If the sinusoids start at the very minimum of the wave, the value of the phase is π radians whereas starting from the top of the wave it will have a phase of *zero*. When two sounds do not have the same phase, it is possible to perceive the difference in the time scale since one of the two is delayed compared to the other. When comparing two signals, there is the need to obtain a *"phase-neutral"* that means the comparison is made taking only into account Amplitude and Frequency. This method is called **autocorrelation** of the signals.

3.1.2 Spectrograms

A **Spectrogram** is the visual representation of an acoustic signal [32]. Basically, a Fourier Transformation is applied to the sound, in such a way to obtain the set of waveforms extracted form the original signal and separate their frequencies and amplitudes. The result is typically depicted in a graph with degrees of amplitude with a *light-dark* representation. Since amplitude represents the *energy*, having a darker shade means that the energy is more intense in a certain range of frequencies - lighter when there is low energy. In 2.10 there is an example of the spectrogram.
The visual feedback of the spectrogram is highly dependent from the **window size** of the Fourier Analysis. In fact, different sizes affect the levels of frequencies and time resolution.
If the window size is *short*, the adjacent **harmonics** are distorted but the time resolution is better [32]. An harmonic is *"an integer multiple of the fundamental frequency"*[33] or component frequencies. This is helpful when we are looking for the *formant structure* because the striations created by the spectrogram highlights the individual pitch periods.
On the other hand, a *wider* window size, helps to locate the harmonics because the band of the spectrogram are narrower.

3.2 Fourier Analysis

Fourier Analysis is the process that decompose a periodic waveform into a set of sinusoids having different amplitudes, phases and frequencies. Yet, if we add those waveforms again, we will obtain the original signal. The analysis has been involved in many scientific applications and the reason is due to the following transform properties:

- Linear transformation - the relationship between two modules is kept
- Exponential function are eigenfunctions of differentiation [34]
- Invertible - derived from the linear relationship

[2]In general, a unit of time is considered a single second

In signal processing, the Fourier analysis is used to isolate singular components of a complex waveform. A set of techniques consist in using **Fourier Transformation** on a signal in such a way to be able to manipulate the data in the easiest way possible but at the same time we have to be capable of inverting the transformation [35] [36]. In the next subsections we describe the fundamental steps for manipulating a signal.

Sampling

Sampling is the process that transform a continuous signal in to a discrete one. Each sample can be either be a single value or a set of values at a certain point in time [6].
Consider a sound signal that varies in time a continuous function $s(t)$. For every T seconds, we need to measure the value of the function. This frame of time is called the *sampling intervar* [37]. To calculate the sequence a sampled function is given as follow: $s(nT), \forall$ integer values of n. Thus, the *sampling rate* is the average number of samples obtained in a range of $T = 1sec$ [6]. An example of sampling is shown in 3.2.

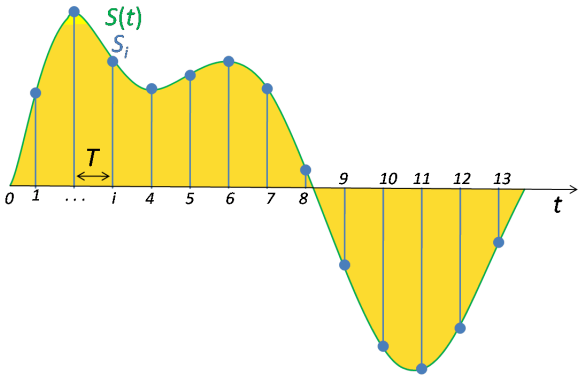

Figure 3.2: Example of signal sampling. The green line represents the continuous signal whereas the samples are represented by the blu lines [6]

As we mentioned above, using the Fourier Analysis we need to be able to reconstruct the original signal from the transformed one. To be able to, the **Nyquist-Shannon** theorem states that the sampling rate has to be larger as twice as the maximum frequency of the signal, in order to rebuild the original signal [38].
The *Nyquist sampling rate* is defined by the following equation:

$$f_s > f_{Nyquist} = 2f_{max} \tag{3.1}$$

Quantization

To finalize the transformation from a continuous signal to a discrete one, we need to *quantized* the signal in such a way to obtain a finite set of values. Unlike sampling in which permits to reconstruct the original signal, quantization is an irreversible operation that introduce a loss of information.
Consider x be the sampled signal and x_q the quantized one where x_q can be expressed as the signal x plus the error e_q. From here we have:

$$x_q = x + eq \Leftrightarrow e_q = x - x_q \tag{3.2}$$

Given the equation above, we can restrict the range of error to $-q/2... + q/2$ because we will not make a larger error than the half of the quantization step. From a mathematical point of view, the error-signal is a random signal with an uniform probability distribution between the range of $q/2and + q/2$, giving the following [39]:

$$p(e) = \begin{Bmatrix} \frac{1}{q} & for \frac{-q}{2} \leq e < \frac{q}{2} \\ 0 & otherwise \end{Bmatrix} \tag{3.3}$$

Given this reason, the quantization error also called quantization noise.

Windowing Signals

Speech sound is a **non-stationary** signal where its properties (amplitude, frequency and pitch) rapidly change over time [40]. Due to the quick changes of those properties, it makes hard to use *autocorrelation* or *Discrete Fourier Transformation*. In 2 we highlighted the fact that phonemes have some invariant properties for a small period of time. Having said that, it is possible to apply methods that will take *short windows* (pieces of signal) and process them. This window is also called **frame**. Typically, the shape of this window is *rectangular* because one of the most used methods are the *Hanning* and *Hamming* in which the window covers the whole amplitude spectrum between a range. In 3.3 there is an example on how the Hamming window is taken from a signal. The rectangle called *Time Record*, is the frame that is extracted and processed by the windowing function.

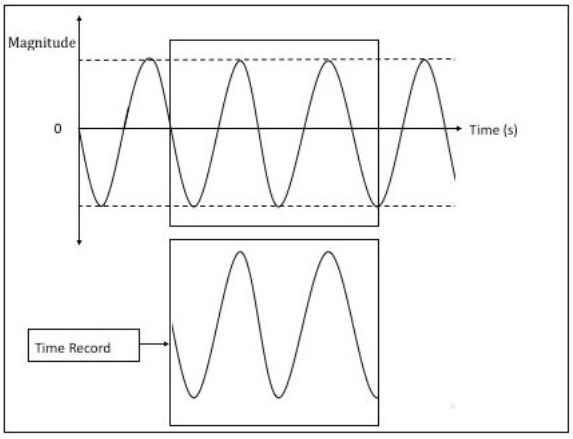

Figure 3.3: Hamming window example on a sinusoid signal

Hann Function

This is one of the most used windowing method in signal processing. The function is discrete and it is defined by 3.4a. The method is a linear combination of the *rectangular function* defined by 3.4b. Starting from the *Euler's formula*, it is possible to inject the rectangular equation as in 3.4c. From here, given the properties of the *Fourier Transformation*, the spectrum of the window function is defined as in 3.4d. Combining the spectrum with 3.4b we obtain 3.4e in which the signal modulation factor *disappears* when the windows are moved around time 0.

$$w(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{3.4a}$$

$$w_r = \mathbf{1}_{[0,N-1]} \tag{3.4b}$$

$$w(n) = \frac{1}{2}w_r(n) - \frac{1}{4}e^{\mathrm{i}2\pi\frac{n}{N-1}}w_r(n) - \frac{1}{4}e^{-\mathrm{i}2\pi\frac{n}{N-1}}w_r(n) \tag{3.4c}$$

$$\hat{w}(\omega) = \frac{1}{2}\hat{w}_r(\omega) - \frac{1}{4}\hat{w}_r\left(\omega + \frac{2\pi}{N-1}\right) - \frac{1}{4}\hat{w}_r\left(\omega - \frac{2\pi}{N-1}\right) \tag{3.4d}$$

$$\hat{w}_r(\omega) = e^{-\mathrm{i}\omega\frac{N-1}{2}}\frac{\sin(N\omega/2)}{\sin(\omega/2)} \tag{3.4e}$$

The reason why this windowing method is one of the most diffuse is due to the *low aliasing*

Zero Crossing Rate

Zero crossing is the point of the function where the sign changes from a positive value to a negative one or vice versa. The method of counting the zero crossings is widely used in speech recognition for estimating the *fundamental* frequency of the signal. The zero-crossing rate is the rate of this positive-negative changes. Formally, it is defined as follow:

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1}\begin{Bmatrix}1 & s_t s_{t-1} < 0\\ 0 & otherwise\end{Bmatrix} \tag{3.5}$$

where s is the signal of length T.

The Discrete Fourier Transform

Before to jump into the definition of the Discrete Fourier Transformation (DFT), we need to introduce the Fourier Transformation (FT) from the mathematical point of view. The FT of a continuous-signal $x(t)$ is defined by the following equation:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt, \qquad \omega \in (-\infty, \infty) \tag{3.6}$$

The discrete operation allows us to transform the equation above from an infinite space in a finite sum as follows:

$$X(\omega_k) = \sum_{n=0}^{N-1} x(t_n) e^{-j\omega_k t_n}, \qquad k = 0, 1, 2, \ldots, N-1 \tag{3.7}$$

where $x(t_n)$ is the *amplitude* of the signal at time t_n (sampling time). T is the sampling period in which the transformation is applied. $X(\omega_k)$ is the *spectrum* of the complex value x at frequency ω_k. Ω is the sampling interval defined by the *Nyquist-Shannon* theorem whereas N is the number of samples.

The motivation behind the DFT is that we want to move the signal from the *Time or space domain* to the *Frequency domain*. This allows us to analyze the spectrum in a simpler way. 3.4 shows the transformation.

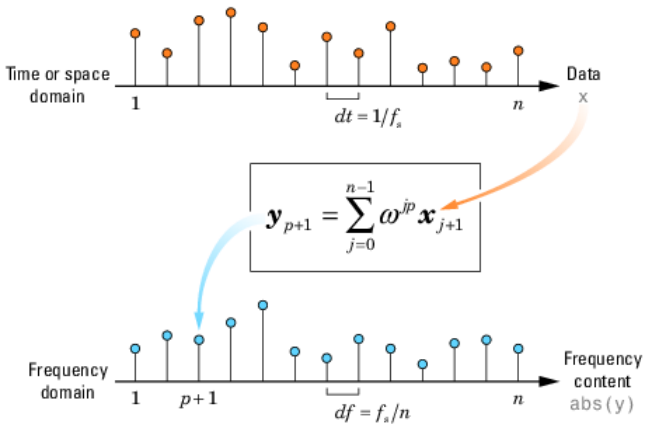

Figure 3.4: DFT transformation [7]

Chapter 4

Speech Recognition

4.1 The Problem of Speech Recognition

4.2 Components

4.3 Naïve Bayes and Gaussian models for classification

naïve

Chapter 5

Artificial Neural Networks

5.1 Artificial Neuron

Our brain is composed by biological neurons and an artificial neuron (or AN) is a representation of it. Every AN can gather signals from other neurons or from the environment, and after an elaboration it transmits another signal to all the other ANs that are connected to it [41]. A rapresentation of AN is depicted in 5.1.
Each connection to the artificial neuron has a numerical weigth associated to it in which the input signal is hold back. The value of the weight can be either positive or negative. In most cases, the sums of each node are weighted and then given as input to a *non-linear* function called **transfer function** or **activation function** [8]. The activate function defines the output value of the node and typically, the *Step Function*, *Sigmoid Function* and a *Softmax Function* are the most used.

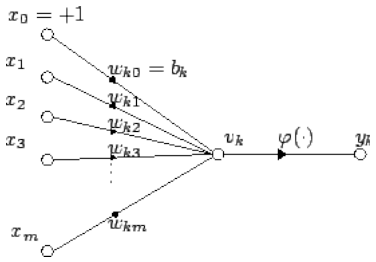

Figure 5.1: Representation of an Artificial Neuron [8]

From the mathematical point of view, we can define an artificial neuron as follow:
given $m + 1$ inputs with signals from x_1 to x_m and weights values from w_0 to w_m. The *bias* is then defined by the input x_0 in which a value of 1 will be assigned. The bias value allows us to *shift* the curve of the activation function to a certain direction and it is defined with $w_{k0} = b_k$ [8].
The output of the AN is:

$$y_k = \varphi\left(\sum_{j=0}^{m} w_{kj} x_j\right)$$

5.2 Network Function

When there are many aritificial neurons interconnected between each other in the different layers, we form a *network*. 5.2 shows an example of ANN where the **inputs** are represented by the first layer in which they send data through the connection to the second group of neuron. The connection between two neuros is called *synapses* where the **weigth** is stored. The second layer is connected to the third one that represents the **output** of the network. There can be multiple stratums between the inputs and the outputs and these are called *hidden layers*.

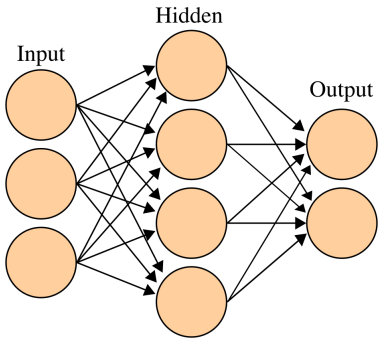

Figure 5.2: Example of ANN [9]

Typically, a neural network is defined by three factors:

1 How the different layers are interconnected

2 How the the weights are updated (learning process)

3 How the neuron's input value is converted to its output activation (activation function)

5.3 Learning

For every application the design of a neural network is different and after it has structured, the NN is ready to be trained. The first step is the initialization of the weigths. Normally they are initialized with random values, although, in [42] they developed a method to *"determining the optimal initial weights of feedforward neural networks based on the Cauchy's inequality and a linear algebraic method"*. More, in [43] they tested seven different weights' initialization for twelve different problems. Thus, even the initialization of the connections values is application-dependant.
The learning paradigms can be grouped in three major categories: **supervised**, **unsupervised** and **reinforcement learning**.

Supervised Learning

This learning technique is the task of inferring a function from labeled training data [44] [45]. The set responsible for training the model is composed by *training examples* in which every sample consists of an *input object* and a **desidered** *output value*. What the algorithms does is to analyze the train dataset and produce an inferred function that will be used to map the new examples [44]. Basically, the algorithm can be seen as *learning* with a *teacher*, in the sense that the there is costantly a feedback on the status of the application.

Unsupervised Learning

While in supervised learning, the system has a desidered output given from the training dataset, in the unsupervised paradigm, the system has to learn to estimate the right output given a new input [46]. In *classification*, this output is a class label whereas in *regression* is a real number. There are several ways to model this kind of learning system. *Clustering*, using *Self-Organizing Maps*, *K-means* or *hierachical clustering* are among the most famous approaches.

Reinforcement Learning

In reinforcement learning, the system takes actions to interact with its own environment. Each action will affect the state of the environment in which will produe a result in a form of either *reward* or *punishment*. The goal of this learning paradigm is to learn which is the best sequence of actions that maximises the rewards or minimises the punishments. There are quite a few approaches to find the best sequence of actions. The most famouse are *Monte Carlo methods*, *Temporal Difference methods* and *Direct Policy search methods*.

5.4 Multilayer Perceptron

The model created by the Multilayer perceptron (MLP) serves to map a set of inputs onto a group of ouputs. The main feature of the MLP is that it a *feedforward* artificial neural network. The MLP is formed by a defined number or layers in which every layer is *fully connected* to following one. Every neuron of the network has a nonlinear activation function, with the exception of the input nodes. The techinque used in the MLP is of the kind of supervised and it uses the **backpropagation** for training the neurons [47]. Backpropagation is discussed more in details in 5.4.

Learning through Backpropagation

The learning phase in the neural network occurs in the moment that the connection weigths change based on the error value in the output. The error is calculated comparing the result the network produced with the expected one. This is a typical example of supervised learning because the network compares the result it just obtained with the one that it was expecting. This process is done through **backproagation**.

In backpropagation, the error output of the node j in the nth sample of the training dataset is given by

$$e_j(n) = d_j(n) - y_j(n) \tag{5.1}$$

where d is the expected output whereas y is the output value obtained from the neuron. The weights are then adjusted in such a way that the error is minimized. With 5.2 we are able to determine the corrections to apply given an output value produced by a perceptron.

$$\varepsilon(n) = \frac{1}{2}\sum_j e_j^2(n) \tag{5.2}$$

At this point using 5.3 we are able to determine the amount of change for each weight. This is done by **gradient descent** which is a first-order optimization algorithm. Gradient descent is used to find *local minima* of a function, where *"it takes steps proportionally to the negative of the gradient of the function in that point"*[48]. The opposite instead, it means that it is approaching to the *local maxima* of the function. ALthough, in this way, the process would be called *gradient ascent*[48].

$$\Delta w_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_j(n) \tag{5.3}$$

In 5.3, η represents the *learning rate* whereas y_i is the output of the previous neuron [47]. The learning rate parameter is one of the most importan parameters when design a neural network. The reason is that, the value used for it ensures that the weights are converging as fast as possible avoiding waivings.

The calculation of the derivates depends on the field v_j where this value changes itself. Continuing, 5.3 can be simplified to 5.4 where ϕ' represents the *derivate* of the activation function described before. Note that this does not changes itself.

$$\frac{\partial \varepsilon(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n)) \tag{5.4}$$

5.5 Deep Learning

Deep Learning has several definitions in which those have been changing in last 10 years. Definition number 5 reported in [49] says the following: *"Deep Learning is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial Intelligence. Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text."*.
There are two main aspects of the level of representations described above:

1) a model consists of several layers of nonlinear processing units

2) for supervised and unsupervised approaches the feature representation have more abstract layers

It is possible to classify Deep Learning as an intersection between different research areas, such as: neural networks, pattern recognition, signal processing, etc..[49]

5.5.1 Deep Neural Networs

As described in 5.5, a *deep neural network* is composed by several hidden layers of perceptrons between the input(s) and the output(s) [50]. As for the artificial neural networks, DNNs are able to model complex non-linear relationships. [51]
Adding extra layers to a ANN permits to each layer to *specialize* in solving a certain problem. For example, in *visual pattern recognition* the neurons in the first layer can try to learn how to recognize edges in a picture, whereas those in the second layer might focus in learning ho to recognize more complex shapes like a square or a triangle built up from the previous edges. The next layers will try to learn even more complex shapes. In the case of speech recognition, we could split the problem in different parts as well. The first layer could focus in recognizing phonemes, the second layer can learn about the pitch voice, and so on.
The usage of multiple hidden layers gives to DNNs some advantages in learning how to solve complex problems compared to the shallow networks.[10] An example of DNN is depicted in 5.3.

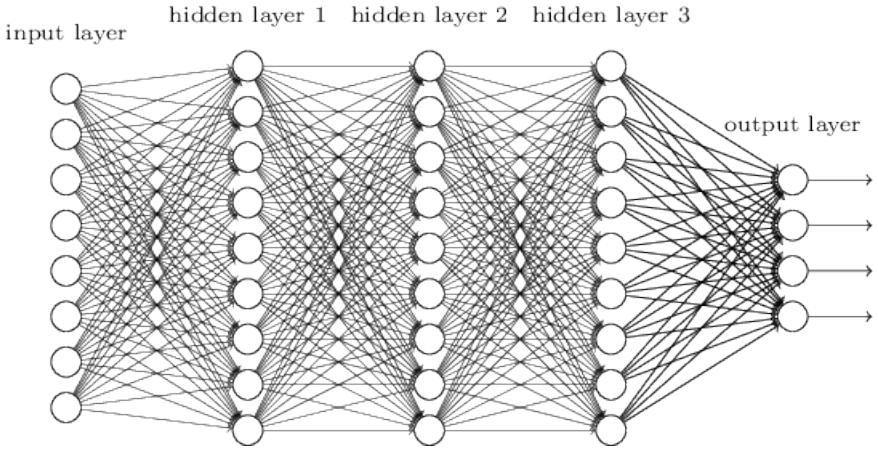

Figure 5.3: Example of Deep Neural Network [10]

Architecture

Deep Neural Networks are oftenly designed as feedforward networks and trained with the *back propagation* algorithm. Although, to train acoustinc models for Automatic Speech Recognition (ASR) *Convolutional Deep Neural Networks* (CNNs) presented better results than the typical feedforward design. Despite the similarity between the two networks, in CNNs *"the neurons in the inputs of hidden units in layer m are from a subset of units in layer m-1, units that have spatially contiguous receptive fields"*. [52]
To update the weights the *stochastic gradient descent* is used as follow:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{5.5}$$

where η is the *learning rate* whereas C is the *cost function*. The activation function and the learning type have a direct influence in the choice of the cost function. For instance, a typical choice of the activation function for a supervised learning approach on a problem of multiclass classification are either **softmax function** or **cross entropy function**. The mathematical definition of *softmax function* is defined in 5.6 where p_j represents the class probability whereas x_k and x_j are the total input to neuron x and j of that layer. *Cross entropy function* is defined by 5.7 where d_j is the target probability of the output neurons j and p_j is the probability output for j after applying the activation function. [51]

$$p_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \tag{5.6}$$

$$C = -\sum_j d_j log(p_j) \tag{5.7}$$

Chapter 6

Implementation

6.1 Data Collection

6.2 PRAAT

6.3 Training the Neural Network

6.4 The Application

6.5 Setting up the Server

Chapter 7

Evaluation

Chapter 8

Conclusions

Chapter 9

Future Works

Bibliography

[1] J. Glass and V. Zue, "6.345 automatic speech recognition," Spring 2003. http://ocw.mit.edu, (Massachusetts Institute of Technology: MIT OpenCourseWare), (Accessed 23 Sep, 2015). License: Creative Commons BY-NC-SA.

[2] "The spectrum of acousting," 2015. accessed 2015-09-28. Available: http://www.hum.uu.nl/uilots/lab/courseware/phonetics/basics_of_acoustics_2/formants.html.

[3] "Rp vowel length: some details," 2015. accessed 2015-10-08. Available: https://notendur.hi.is/peturk/KENNSLA/02/TOP/VowelLength0.html#lengths.

[4] "How do i read a spectrogram ?," 2015. accessed 2015-10-28. Available: https://home.cc.umanitoba.ca/~robh/howto.html.

[5] "Example of autocorrelation," 2015. accessed 2015-10-28. Available: http://www.eng.usf.edu/~lazam2/Project/sht_time_timedom/xmp_acr.htm.

[6] "Sampling (signal processing)," 2015. accessed 2015-11-08. Available: https://en.wikipedia.org/wiki/Sampling_(signal_processing).

[7] "Discrete fourier transform (dft)," 2015. accessed 2015-11-08. Available: http://www.mathworks.com/help/matlab/math/discrete-fourier-transform-dft.html.

[8] "Artificial neuron," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Artificial_neuron.

[9] "Artificial neural network," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Artificial_neural_network.

[10] "Why are deep neural networks hard to train?," 2015. accessed 2015-09-08. Available: http://neuralnetworksanddeeplearning.com/chap5.html.

[11] T. M. Derwing and M. J. Munro, "Second language accent and pronunciation teaching: A research-based approach," *Tesol Quarterly*, pp. 379–397, 2005.

[12] P. Medgyes, "When the teacher is a non-native speaker," *Teaching English as a second or foreign language*, vol. 3, pp. 429–442, 2001.

[13] A. Gilakjani, S. Ahmadi, and M. Ahmadi, "Why is pronunciation so difficult to learn?," *English Language Teaching*, vol. 4, no. 3, p. p74, 2011.

[14] M. Rost and C. Candlin, *Listening in language learning*. Routledge, 2014.

[15] "Word stress - british council," 2015. accessed 2015-09-28. Available: https://www.teachingenglish.org.uk/article/word-stress.

[16] D. Edge, K.-Y. Cheng, M. Whitney, Y. Qian, Z. Yan, and F. Soong, "Tip tap tones: mobile microtraining of mandarin sounds," in *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pp. 427–430, ACM, 2012.

[17] A. Head, Y. Xu, and J. Wang, "Tonewars: Connecting language learners and native speakers through collaborative mobile games," in *Intelligent Tutoring Systems*, pp. 368–377, Springer, 2014.

[18] P. Boersma and D. Weenink, "{P} raat: doing phonetics by computer," 2010.

[19] "Nasal consonants," 2015. accessed 2015-10-28. Available: https://en.wikipedia.org/wiki/Nasal_consonant.

[20] "Dipthong," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Diphthong.

[21] "Schwa," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Schwa.

[22] "What are fricatives ?," 2015. accessed 2015-10-28. Available: http://www.pronuncian.com/Lessons/default.aspx?Lesson=9.

[23] "Stop consonants," 2015. accessed 2015-10-28. Available: https://en.wikipedia.org/wiki/Stop_consonant.

[24] P. Ladefoged and I. Maddieson, "The sounds of the world's languages," *Language*, vol. 74, no. 2, pp. 374–376, 1998.

[25] "Maxillary lateral incisor," 2015. accessed 2015-10-28. Available: https://en.wikipedia.org/wiki/Maxillary_lateral_incisor.

[26] "Syllable, stress & accent," 2015. accessed 2015-10-28. Available: http://hubblesite.org/reference_desk/faq/answer.php.id=73&cat=light.

[27] P. Roach and E. Phonetics, "Phonology: A practical course," *Cambridge UP Cambridge*, 2000.

[28] "What is the relationship between wavelength, frequency and energy?," 2015. accessed 2015-10-28. Available: http://www.personal.rdg.ac.uk/~llsroach/phon2/mitko/syllable.htm.

[29] J. Laver, *Principles of phonetics*. Cambridge University Press, 1994.

[30] S. J. Orfanidis, *Introduction to signal processing*. Prentice-Hall, Inc., 1995.

[31] "Properties of sinusoids," 2015. accessed 2015-10-28. Available: http://web.science.mq.edu.au/~cassidy/comp449/html/ch03s02.html.

[32] "So what is a spectrogram anyway?," 2015. accessed 2015-11-08. Available: https://home.cc.umanitoba.ca/~robh/howto.html.

[33] "Harmonic," 2015. accessed 2015-11-08. Available: https://en.wikipedia.org/wiki/Harmonic.

[34] L. C. Evans, "Partial differential equations and monge-kantorovich mass transfer," *Current developments in mathematics*, pp. 65–126, 1997.

[35] "Fourier analysis," 2015. accessed 2015-11-08. Available: https://en.wikipedia.org/wiki/Fourier_analysis#CITEREFEvans1998.

[36] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.*, vol. 1, 1975.

[37] M. Weik, *Communications standard dictionary*. Springer Science & Business Media, 2012.

[38] "Sampling and quantization," 2015. accessed 2015-11-08. Available: https://courses.engr.illinois.edu/ece110/content/courseNotes/files/?samplingAndQuantization.

[39] "Digital signals - sampling and quantization," 2015. accessed 2015-11-08. Available: http://rs-met.com/documents/tutorials/DigitalSignals.pdf.

[40] "Windowing signal processing," 2015. accessed 2015-11-08. Available: http://www.cs.tut.fi/kurssit/SGN-4010/ikkunointi_en.pdf.

[41] A. P. Engelbrecht, *Computational intelligence: an introduction*. John Wiley & Sons, 2007.

[42] J. Y. Yam and T. W. Chow, "A weight initialization method for improving training speed in feedforward neural network," *Neurocomputing*, vol. 30, no. 1, pp. 219–232, 2000.

[43] M. Fernandez-Redondo and C. Hernandez-Espinosa, "Weight initialization methods for multilayer feedforward.," in *ESANN*, pp. 119–124, 2001.

[44] "Supervise learning," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Supervised_learning.

[45] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[46] Z. Ghahramani, "Unsupervised learning," in *Advanced Lectures on Machine Learning*, pp. 72–112, Springer, 2004.

[47] "Multilayer perceptron (mlp)," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron.

[48] "Gradient descent," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Gradient_descent.

[49] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[50] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[51] "Deep learning," 2015. accessed 2015-09-08. Available: https://en.wikipedia.org/wiki/Deep_learning.

[52] "Convolutional neural networks (lenet)," 2015. accessed 2015-09-08. Available: http://deeplearning.net/tutorial/lenet.html.