ON THE ANALYSIS OF FRENCH PHONETIC IDIOSYNCRASIES FOR ACCENT RECOGNITION

A PREPRINT

Pierre Berjon
INP-ENSEEIHT
Toulouse, France

Avishek Nag
University College Dublin
Dublin, Ireland

Philippe Brodeur
Overcast
Dublin/United States

Marianne Checkley
Camara Education
Dublin, Ireland

Annette Klinkert
European Science Engagement Association
Vienna, Austria

Soumyabrata Dev
University of Dublin
Dublin, Ireland

September 30, 2020

ABSTRACT

Since the speech recognition system has been created, it has developed significantly, but it still has a lot of problems. As you know, any specific natural language may own at least one accent. Despite the identical word phonemic composition, if it is pronounced in different accents, as a result, we will have sound waves, which are different from each other. Differences in pronunciation, in accent and intonation of speech in general, create one of the most common problems of speech recognition. If there are a lot of accents in language we should create the acoustic model for each separately. Here, we will try to do it with the French accent.

***K*eywords** Accent Recognition · French Accent Classification

1 Introduction

Accent recognition is one of the most important topics in automatic speaker and speaker-independent speech recognition (SI-ASR) systems in recent years. The growth of voice-controlled technologies has becoming part of our daily life, nevertheless variability in speech makes these spoken language technologies relatively difficult. One of the profound variability is accent. By classifying accent types, different models could be developed to handle SI-ASR.

Dialect/accent refers to different ways of pronouncing/speaking a language within a community. Examples could be American English vs. British English speakers or the Spanish speakers in Spain vs. Caribbean. During the past few years, there have been significant attempt to automatically recognize the dialect or accent of a speaker given his or her speech utterance. Recognition of dialects or accents of speakers prior to automatic speech recognition (ASR) helps in improving performance of the ASR systems by adapting the ASR acoustic and/or language models appropriately. Moreover, in applications such as smart assistants as the ones used in smartphones, by recognizing the accent of the caller and then connecting the caller to agent with similar dialect or accent will produce more user friendly environment for the users of the application. However, the main deep learning models do not allow the creation of an accent recognition system whose accuracy exceeds 0.7. We obtained that score in using three kinds of Machine Learning and Deep Learning methods, which will be described later.

One of the reasons we are having trouble to have a good accuracy in the accent recognition problem is the lack of knowledge we have of English syllabic structure. Indeed, in order to approximate English phonology, we have to understand the native language similarities of articulation, intonation, and rhythm. In the past, the research has focused on phone inventories and sequences, acoustic realizations, and intonation patterns. That's why we have to study the English syllable structure.

The main problematic of the word recognition is the understanding of the syllable. It usually consists of an obligatory vowel with optional initial and final consonants. One familiar way of subdividing a syllable is into Onset and Rhyme. All syllables in all languages consist of Onset and Rhyme (phonetically, at least). However, these categories alone do not indicate where the syllable is placed within the word. In order to capture foreign accents in English, we want to highlight those constituents of the syllable that are most likely to prove difficult for speakers of languages in which they are not contained [1].

We separate the syllables into 3 parts:

1. **Proclitic** : Syllable component that begins with a morpheme and doesn't have any other, like "s" in "still" or"shrugged".
2. **Core** : Syllable component with an obligatory vowel.
3. **Enclitic** : Syllable component that ends with a morpheme finally and doesn't have any other.

These three categories capture a syllable structure for English language.

Only some languages have Proclitics and Enclitics. In contrast to English, tone languages use tone for the same function. Syllable structures in tone languages tend to be comparatively simple in terms of phone segments, but are complicated by the extension of a tone for the duration of a syllable or syllables expressing a grammatical unit, usually the word. This difference in language typology has a strong effect on the ability to pronounce English in parts of the syllable that demarcate grammatical units. In [1], they chose to use a Vietnamese speech dataset instead of a Lebanese or an Arabic one, because these ones have much more in common with English (it would be harder to recognize a Lebanese accent than a Vietnamese accent) [1].

Thus, we will focus on solving this problem by trying to understand what is lacking in the models used by studying the structure of the languages used.

Here, we will focus on the specifications of the French language, a non tonal language. Indeed, the main objectives of this research are to find which are the idiosyncrasies of French people that lead a model into predicting the wrong accent.

1.1 Related work

In order to do this research, I have for now read four main papers. The first one, *SCoPE, Syllable Core And Periphery Evaluation: Automatic Syllabification And Application To Foreign Accent Identification*, gave me everything I needed to know about the Proclitics and Enclitics I talked about before. The second one, *Solving the Problem of the Accents for Speech Recognition Systems*, they have developed an approach, which is used to solve above mentioned problems and create more effective, improved speech recognition system of Georgian language and of languages, which are similar to Georgian language. The third one, *Automatic Detection of Foreign Accent for Automatic Speech Recognition*, an automatic method of detection of the degree of foreign accent is proposed and results are compared with accent labeling carried out by an expert phonetician. In *On the Modelization of Allophones in an HMM Based Speech Recognition System*, they give a new approach for modelling allophones in a speech recognition system based on hidden Markov models.

1.2 Contributions of the paper

The main contributions of this paper[1] can be summarized as follows:

- highlighting the problem of the limit in the context of the study of Accent Recognition
- highlighting French idiosyncrasies restricting the precision of Deep Learning models; and
- highlighting the incidence of these idiosyncrasies in the spectrograms, and therefore the models in question.

The rest of the paper is structured as follows. Section 2 discusses the data and the methods we used in our preliminary study (dataset and neural networks) and Section 3 discusses results we obtained with these methods. In Sections 4 and 5, we analysed the French speakers idiosyncrasies and their consequences on spectrograms.

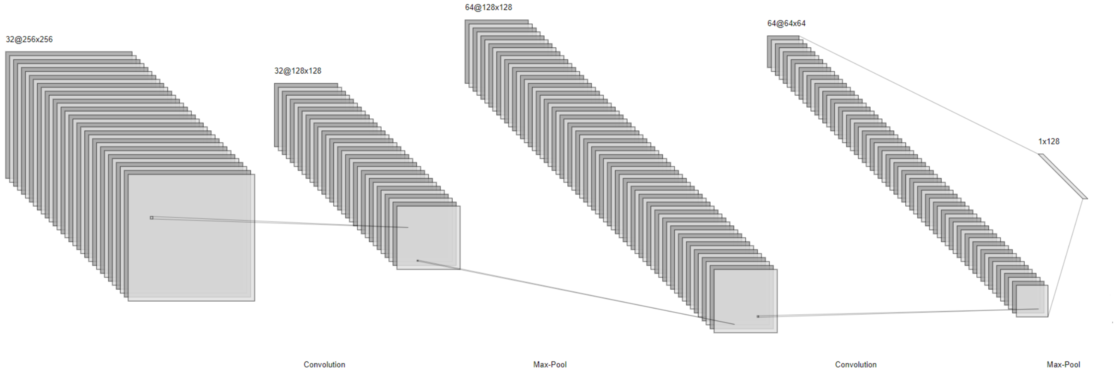
[1] With the spirit of reproducible research, the code to reproduce the results in this paper is shared at `https://github.com/pberjon/Article-Accent-Recognition`.

2 Data & Methods

2.1 Features and approaches

Spectrograms are pictorial representation of sound. The x-axis represents time in seconds while the y-axis represents frequency in Hertz. Different colors represent the different magnitude of frequency at a particular time. We can think of the spectrogram as an image. Once the audio file is converted to an image, the problem reduces to an image classification task. Based on the number of images, algorithms like Support Vector Machines(SVM), etc. are used to classify sound, validate the speaker, speaker diarisation, etc.

We used different Machine Learning and Deep Learning models, and the first one is a two convolutional layers neural network with 5 different accents. This neural network is a 2-layer Convolutional Neural Network : one with 32 filters and a relu activation function, and another one with 64 filters and a relu activation function :



We will focus on this one for the rest of our work, like we will discuss it later.

2.2 Dataset

Everyone who speaks a language, speaks it with an accent. A particular accent essentially reflects a person's linguistic background. When people listen to someone speak with a different accent from their own, they notice the difference, and they may even make certain biased social judgments about the speaker.

In this research, we used the Speech Accent Archive. It has been established to uniformly exhibit a large set of speech accents from a variety of language backgrounds. Native and non-native speakers of English all read the same English paragraph and are carefully recorded.

This dataset allows us to compare the demographic and linguistic backgrounds of the speakers in order to determine which variables are key predictors of each accent. The speech accent archive demonstrates that accents are systematic rather than merely mistaken speech.

It contains 2140 speech samples, each from a different talker reading the same reading passage. Talkers come from 177 countries and have 214 different native languages. Each talker is speaking in English. The samples were collected by many individuals under the supervision of Steven H. Weinberger, the most up-to-date version of the archive is hosted by George Mason University and can be found here : `https://www.kaggle.com/rtatman/speech-accent-archive`.

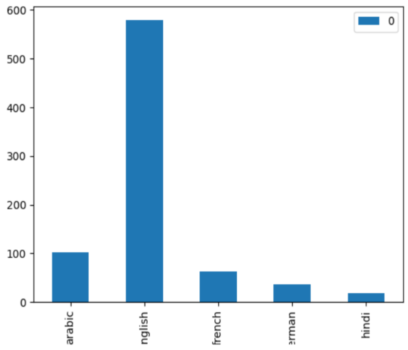


Figure 1: Samples repartition in the dataset

3 Results and Discussion

3.1 Accuracy rate

SVM	CNN with 2 layers	CNN with 4 layers
0.35	0.70	0.65

With regular Machine Learning methods as SVM, we obtained low accuracies (around 0.3 and 0.4); the impact of Deep Learning methods is quite clear here : as soon as we use Convolutional Neural Networks, the accuracy takes at least the value of 0.6. However, we can see that we won't obtain an optimal score if we use too many layers in our model. Depending upon how large our dataset is, the CNN architecture is implemented. Adding layers unnecessarily to any CNN will increase our number of parameters only for the smaller dataset. It's true for some reasons that on adding more hidden layers, it will give a better accuracy. That's true for larger datasets, as more layers with less stride factor will extract more features for the input data. In CNN, how we play with the architecture is completely dependent on what our requirement is and how our data is. Increasing unnecessary parameters will only overfit your network, and that's the reason why our CNN with 2 layers has better results than with 4.

3.2 Multi-class classification

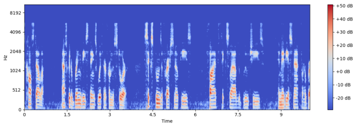
We obtained these results in the confusion matrixes with the 2-layer CNN and the SVM method :

SVM					2-layer CNN				
Classes	**ACC**	**AGF**	**AUC**	**GI**	**Classes**	**ACC**	**AGF**	**AUC**	**GI**
English	0.42391	0.21774	0.36781	-0.26437	English	1.0	1.0	1.0	1.0
Arabic	0.71739	0.0	0.5	0.0	Arabic	0.95	0.71	0.74	0.48
French	0.34783	0.51315	0.49171	-0.01658	French	0.85	0.84	0.84	0.69
German	0.92391	0.0	0.5	0.0	German	0.84	0.80	0.80	0.61
Hindi	0.95652	0.0	0.5	-0.01124	Hindi	0.87	0.32	0.53	0.06

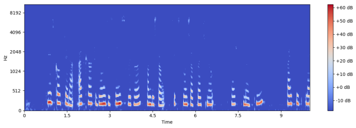
Here, we can see that the "classical" Machine Learning methods are quite ineffective and that the Deep Learning methods stand out clearly in accent recognition; that's why we will use the 2-layer CNNs as a reference for the rest of the paper (and in future works).

In most case, the SVM method is not powerful enough for us to have a good accuracy. That can be explained with the results we had on the Gini Index : all the values taken by the index are quite low (negative values are actually quite low positive values), which means that, regarding to the SVM analysis, the spectrograms are quite similar; the method is not selective enough to clearly determine the accent (which is also shown by the AGF values). However, the SVM method is not totally to be excluded: in the context of the Hindi accent or the German accent, the SVM turns out to be more effective than all the Deep Learning methods used.

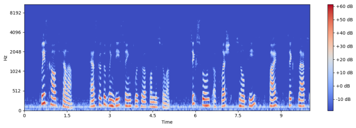
Here are a few examples of spectrograms for the languages we used in the model :



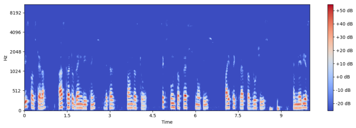
(a) English sample spectrogram



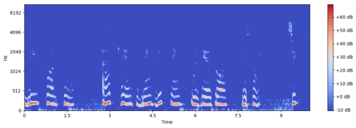
(b) French sample spectrogram



(c) German sample spectrogram



(d) Hindi sample spectrogram



(e) Arabic sample spectrogram

4 French speakers idiosyncrasies

4.1 French-infused vowels

Nearly every English vowel is affected by the French accent. French has no diphthongs, so vowels are always shorter than their English counterparts. The long A, O, and U sounds in English, as in say, so, and Sue, are pronounced by French speakers like their similar but un-diphthonged French equivalents, as in the French words sais, seau, and sou. For example, English speakers pronounce say as [seI], with a diphthong made up of a long "a" sound followed by a sort of "y" sound. But French speakers will say [se] - no diphthong, no "y" sound.

English vowel sounds which do not have close French equivalents are systematically replaced by other sounds:

With an English accent	With a French accent
short A, as in fat	pronounced "ah" as in father
long A followed by a consonant, as in gate	pronounced like the short e in get
ER at the end of a word, as in water	pronounced air
short I, as in sip	pronounced "ee" as in seep
long I, as in kite	elongated and almost turned into two syllables: [ka it]
short O, as in cot	pronounced either "uh" as in cut, or "oh" as in coat
U in words like full	pronounced "oo" as in fool

4.2 Dropped Vowels, Syllabification, and Word Stress

French people pronounce all schwas (unstressed vowels). Native English speakers tend toward "r'mind'r," but French speakers say "ree-ma-een-dair." They will pronounce amazes "ah-may-zez," with the final e fully stressed, unlike native speakers who will gloss over it: "amaz's." And the French often emphasize the -ed at the end of a verb, even if that means adding a syllable: amazed becomes "ah-may-zed."

Short words that native English speakers tend to skim over or swallow will always be carefully pronounced by French speakers. The latter will say "peanoot boo-tair and jelly," whereas native English speakers opt for pean't butt'r 'n' jelly.

Because French has no word stress (all syllables are pronounced with the same emphasis), French speakers have a hard time with stressed syllables in English, and will usually pronounce everything at the same stress, like actually, which becomes "ahk chew ah lee." Or they might stress the last syllable - particularly in words with more than two: computer is often said "com-pu-TAIR."

4.3 French-accented Consonants

H is always silent in French, so the French will pronounce happy as "appy.". Once in a while, they might make a particular effort, usually resulting in an overly forceful H sound - even with words like hour and honest, in which the H is silent in English.

J is likely to be pronounced "zh" like the G in massage.

R will be pronounced either as in French or as a tricky sound somewhere between W and L. Interestingly, if a word starting with a vowel has an R in the middle, some French speakers will mistakenly add an (overly forceful) English H in front of it. For example, arm might be pronounced "hahrm."

TH's pronunciation will vary, depending on how it's supposed to be pronounced in English: - voiced TH [ð] is pronounced Z or DZ: "this" becomes "zees" or "dzees" - unvoiced TH is pronounced S or T: "thin" turns into "seen" or "teen"

Letters that should be silent at the beginning and end of words (psychology, lamb) are often pronounced.

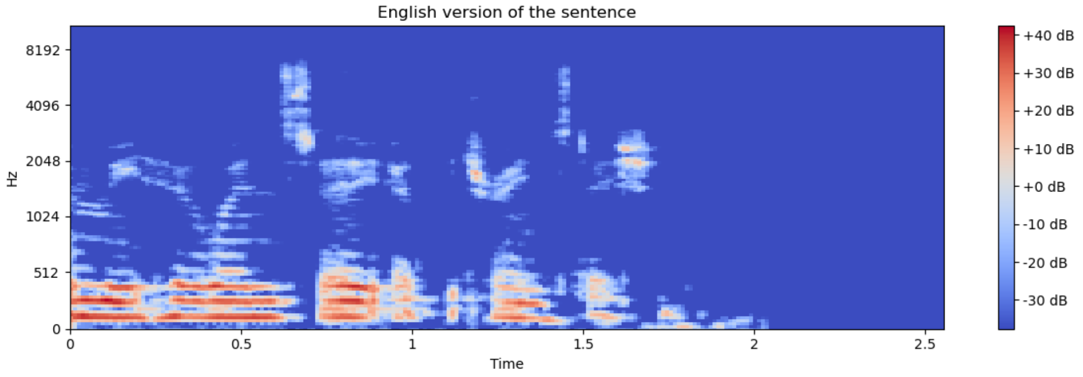
5 Idiosyncrasies Consequences on Spectrograms

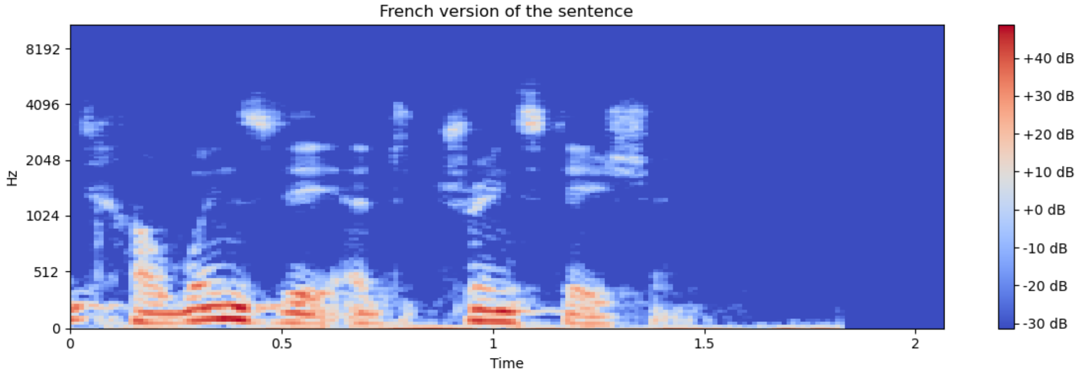
We will now study the idiosyncrasies of the French language.

5.0.1 The un-diphthonged "y"

Firstly, we will analyze differences on the spectrograms for the word "Wednesday", where the French speaker is not supposed to use the "y" sound, like it was explained in French-infused Vowels.

Here are the spectrograms of an English speaker and a French speaker of the sentence “and we will go meet her Wednesday at the train station” :





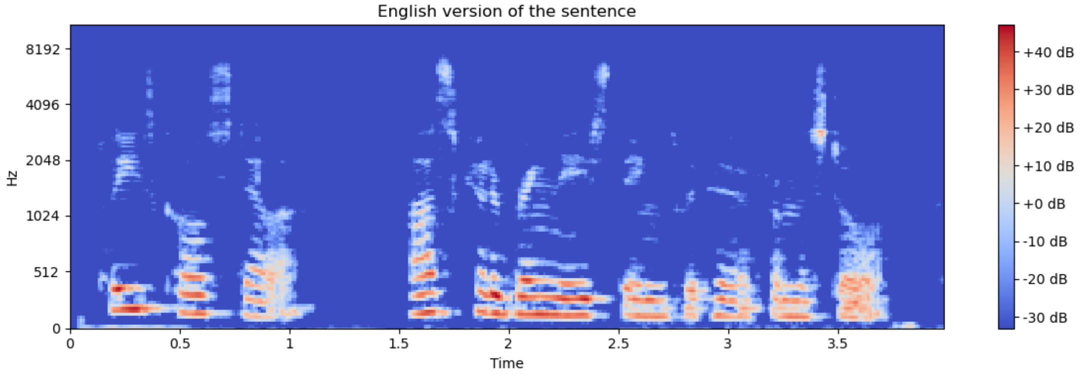
“Wednesday” in English version : 0.8s-1.4s.

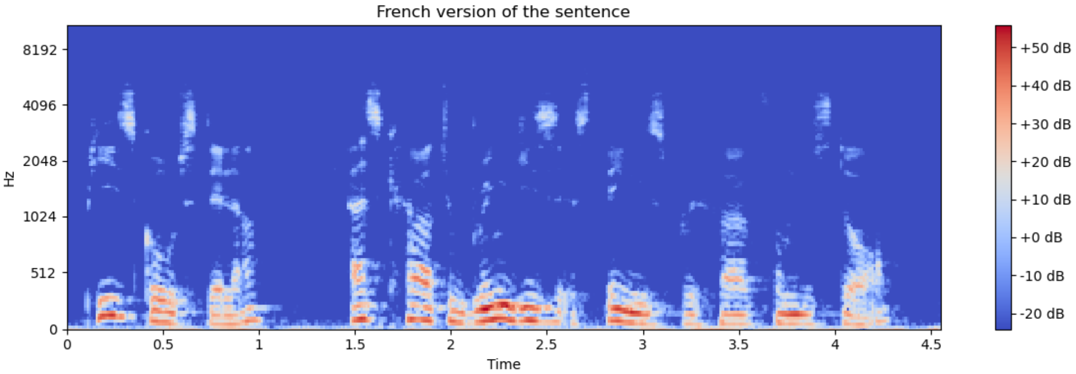
“Wednesday” in French version : 0.7s-1.10s.

We can see, as expected, that at the end of the word (1.3-1.4 for English and 1.05-1.1 for French), the “y” is almost not even pronounced by the French speaker, while the English speaker pronounced it clearly.

5.0.2 Voiced TH [ð] is pronounced Z or DZ

French people tend to say "zees" instead of “these”.That’s what we can see in the sentence “Please call Stella, ask her to bring these things from the store.”.





It's quite complicated to delimit the word "these" in this sentence because it is quite quick, so we will delimit "bring these", as the word "bring" does not represent a major problem for French speakers.

"Bring these" in English version : 2.5s-3s.

"Bring these" in French version : 2.6s-3.2s.

Here, we see that french people tend to diminish the importance of the word "bring" but accentuate the word "these", whereas English speakers seem to pronounce the sequence "bring these" at the same frequency. we think that's why, for French speakers, the "th" sounds like "z". Indeed, the closest sound to "th" is "z" in the French language, so it's only natural for us to use it. Nevertheless, we think the reason why they accentuate it (because we could just use the sound "z" more discreetly) is because of the role of words like "these", "the", "this"... They're articles, and in the French language, they tend to accentuate the most important parts of the sentence, which made this French speaker diminish "bring", and accentuate "these".

Thus, French speakers idiosyncrasies have a direct impact on audio samples spectrograms. Then, we can easily understand why these idiosyncrasies have a direct impact on the results of Deep-Learning models : the first reason why we use spectrograms in order to develop Speech Recognition Systems is to turn an audio classification problem into an image classification problem. Then, if the idiosyncrasies of a specific language have that much effect on spectrograms, it's quite logical that our Accent Recognition Systems' accuracy is limited.

6 Conclusion and Future work

we have argued that the classical Deep-Learning models are not powerful enough to clearly predict someone's accent. Thus, we decided to study the differences between tonal and non-tonal languages, in order to clearly identify the obstacles that prevent us from achieving better results in Accent Recognition. To fulfill that purpose, we decided to begin the analysis with the French accent, which is a non-tonal language, by the study of French speakers idiosyncrasies : the characteristics of the spoken French language having a direct impact on the way French people pronounce English words. In addition, we determined the consequences these idiosyncrasies have on spectrograms, consequently on Deep-Learning models.

In future works, we would like to work further on the subject of French idiosyncrasies, by building a model which determines if an idiosyncrasy is present in an audio sample or not : this would allow us to more easily determine the presence of a French accent in an audio sample.

References

[1] Kay Berkling, Julie Vonwiller, Chris Cleirigh. SCoPE, Syllable Core And Periphery Evaluation: Automatic Syllabification And Application To Foreign Accent Identification.

[2] Irakli Kardava, Jemal Antidze and Nana Gulua. Solving the Problem of the Accents for Speech Recognition Systems.

[3] Bartkova Katarina and Denis Jouvet. Automatic Detection of Foreign Accent For Automatic Speech Recognition.

[4] K. Bartkova, D. Jouvet and J. Monné. On the Modelization of Allophones in an HMM Based Speech Recognition System.

[5] Weinberger, S. (2013). Speech accent archive. George Mason University. This datasets is distributed under a CC BY-NC-SA 2.0 license.