UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

RAMBLE: robust acoustic modeling for Brazilian learners of English

Christopher Dane Shulby

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Christopher Dane Shulby

RAMBLE: robust acoustic modeling for Brazilian learners of English

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
July 2018

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Shulby, Christopher Dane
S562r RAMBLE: robust acoustic modeling for Brazilian learners of English / Christopher Dane Shulby; orientador Sandra Maria Aluísio. -- São Carlos, 2018.
160 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018.

1. Natural Language Processing. 2. Speech Recognition. 3. Acoustic Modeling. 4. Convolutional Neural Networks. 5. Statistical Machine Learning Theory. I. Aluísio, Sandra Maria, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Christopher Dane Shulby

RAMBLE: modelagem acústica robusta para estudantes brasileiros de Inglês

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluísio

USP – São Carlos
Julho de 2018

This Dissertation is dedicated to:
Julia Brabo Shulby
That she may be inspired to dream and become a life-long learner.

ACKNOWLEDGEMENTS

Firstly, I would like to thank all of those who have pushed me to dream and work hard. I will try and name as many as I can here, but you know who you are.

Above all I would like to thank my family for their emotional support as I have worked full-time, became a father and somehow finished this dissertation. First, my mother Janet L. Willett, for always being there for me when I needed the support my entire life. We didn't have things easy but we always made it work and we have become stronger for it. My wife Marina for her love and support, even when she reluctantly gave up countless weekend plans so that I could work on this dissertation. My daughter Julia B. Shulby has been here since less than 50% of the journey but has made 100% of the difference. Being a father to this beautiful little girl has shown me what is most important and what I need to do to one day get her this far, or hopefully farther. My in-laws Maria de Lourdes Brabo Cruz and Tomás Caceres Cruz for their helpful advice, especially as it comes to braving the academic world here in Brazil... which is not as straight forward as one may think.

I certainly need to thank all of those who have inspired me academically throughout the past three decades. In my early years these people were Mr. Youngblood in Greer and Frau Weindel at St. Paulusheim in Bruchsal. Both have given me a great appreciation for Goethe, Schiller and Brecht.

Wendy Cuellar (née Logan) should be credited as the person who most inspired me to become a life-long learner, built blanket forts with me and taught me that dinner wasn't complete unless the dictionary comes out. She also gave me an insatiable appetite for books (I read the entire Tolkien series and Little House on the Prairie series by the fourth grade... I was a weird kid). Grandpa Zack should be credited for much more than his never-ending imagination and story-telling. He has always been a great advice giver, especially on life's tough choices and being the hardest worker in the field, no matter which it is. Also, he is the one who shared his passion for engineering early in my life. Grandma Carol, on the other hand, is the person who always fueled my creativity. Countless hours painting and sketching in charcoal have had a profound impact on the way I look at the world.

The Brüwer and Pabst Familes also deserve the greatest thanks possible for al-

ways being there for me in a time when I was becoming a young man. Wolfgang Brüwer taught me the value of thinking critically about even the smallest things in life. Gisela, who always prepared me for every day and always took the time to turn mistakes into teachable moments. Moritz, Olga and Gloria for their companionship. Axel Pabst is the person who showed me the value and cultural significance of Badisch. This was one of the main factors that pushed me towards linguistics. He also taught me that even when life is dreadfully busy, you can always make time for your family, that is what Sundays or for. Jutta was always my greatest Badisch tutor and sometimes I got lost in her stories but her love is endless and an hour in the kitchen with her can change your whole perspective on life. Helen and Tobi also deserve a special place her for their support and friendship.

In college, without a doubt I have an amazing group of world-class professors to thank.I Would not be where I am today without The Ohio State University.
Dr. Mary Beckman, Dahee Kim (now Dr. Kim), Dr. Beth Hume, Dr. Cynthia Clopper and Dr. Dave Odden infected me with their contagious love for all P-things and the wonderful world of Laboratory Phonetics. Dr. Detmar Meurers, Dr. Kathy Corl, Dr. Chris Brew and Dr. Michal White were the first people to give me an opportunity to work with Computational Linguistics. DJ Hoover deserves a special mention for my ramp up with Machine Learning, PoS tagging and programming at that time. Also a very special thanks to Dr. Brian Joseph, Dr. Julie McGory, Dr. Carl Pollard and Dr. Bob Levine For the many hours of classwork and discussions about an infinite number of topics. Last but not least, the group who most shaped my mentality as a researcher is without a doubt the "Phonies". The countless hours of paper discussions, pre-conference talks and the endless supply of chips and pretzels are probably what made research "real" for me.
On That note OH-IO. Go Bucks and M*ck Fichigan!

My debt that can never be repaid is to the people who made college possible for me. Beyond working full-time, I needed a great deal of help to make that dream true. There are two people who deserve more than words can say and I will remember them always. My late Aunt Susan Willett and my late Grandmother Rosemary A. "Paw" Willett (née Konkel). Without them nothing would have been possible. My Grandmother is also the person who fueled my passion for German since I was a small child. Her stories about life during wartime and the challenges faced by our family of immigrants always fascinated me.

Since 2012 I have called the NILC laboratory home. What started with an informal conversation with Dr. Magali Duran turned into an incredible experience at the greatest laboratory for Natural Language Processing of Portuguese on the planet. For me, the NILC was more than just that. It was and still is a place where Computational Linguis-

tics actually happens. It is a place where people work together to solve highly complex problems by lending their wide variety of skill sets and honing the skills which they are developing. It is the place where I truly became a hybrid computer scientist and linguist. Firstly, I must thank my fellow colleagues who since the beginning received me as a friend and colleague. A special thanks to Pedro Balage, who with Juliana Balage are also the godparents to my daughter Julia. Fernando Asevedo Nobrega, who not only served as the director of the labs coffee (all programs understand the value of this position), but also was one of the most collaborative researchers I have ever known. Nathan Siegle Hartmann and Andre Cunha for all of their help in ramping up my skills as a computer scientist. Without them I would not be the professional I am today. I would like to thank Lucas Avanço for his friendship and the countless study hours fueled almost 100% on Tereré. Erick "Carioca" Fonseca deserves a special place for all of the philosophical linguistics discussion, often followed by creative implementation ideas, some of which may actually work. Alessandro Bokan also should be mentioned here for all of the fun python times and Petrus discussions, as well as good talk about Pisco. Leandro Borges deserves a special mention as he got me out of jams in my codes multiple times. Another person who deserves an enormous amount of thanks is Sara Candeias for her support and interest in my project. While she is not directly involved with the NILC, she is certainly an extension of the NILC in Portugal. She is one of the most brilliant computational linguists for the Portuguese language and her advice throughout the years has always been extremely valuable.

Of course the person with the most direct impact in this dissertation was my advisor, Dr. Sandra Maria Aluísio. She is the dreamer of the NILC and someone who is never afraid to tread the unknown. She has always battled for more collaboration among linguists and computer scientists and has gone to battle for me countless times to make this dissertation possible. She is in every way the "Mother" of this work. I should also thank along with Sandra, who created our speech processing group at the NILC, my two most valued colleagues Vanessa Marquiafável and Gustavo Mendonça. Bouncing ideas and discussions with them was probably the greatest impact early in our careers and it is no wonder how all three of us have gotten where we are today. Gustavo is a leader at Google as I am at Samsung and Vanessa is a fearless entrepreneur. Also, from the ICMC, the professor who made me a true lover of Machine Learning and Statistical Learning Theory is Dr. Rodrigo Mello. Without him and his student, Martha Dais Ferreira, this dissertation would have turned out very differently. I would also like to thank professors Dr. Gustavo Batista for his advice on over-sampling techniques and unbalanced data sets and Dr. Thiago Pardo for his teachings on Natural Language Processing and advice throughout my career at the University of São Paulo. A very special thanks also goes out to Dorly Piske for the help in recruiting "volunteers" for the corpus.

I would like to thank several institutions for their support throughout the years.

The first is the ICMC (Instituto de Ciências Matemáticas e de Computação) at the University of São Paulo. Not only has this incredible institution always provided world-class research and education but have also been helpful to me in every step of the way.

During much of my time as a PhD student I worked at CPqD (Centro de Pesquisa e Desenvolvimento). Beyond the great professional experiences which without a doubt have contributed to my technical expertise, Norberto Alves Ferreira and his platform always made it possible for me to work and study simultaneously.

Finally, I would like to thank SIDI (Samsung Instituto de Desenvolvimento para a Informática). Samsung has always created the incentives necessary for me to study and work as well as provided the platform to present my work for the benefits of all involved. A special thanks to Vitorino Sin, Ivan Brunetto and Chung Lee for their help during this time.

My advice to future data scientists:
"Forget about the fancy tools,
look at the data."

RESUMO

SHULBY, C. D. **RAMBLE: modelagem acústica robusta para estudantes brasileiros de Inglês**. 2018. 160 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Os ganhos obtidos pelas atuais técnicas de aprendizado profundo frequentemente vêm com o preço do big data e nas pesquisas em que esses grandes volumes de dados não estão disponíveis, uma nova solução deve ser encontrada. Esse é o caso do discurso marcado e com forte pronúncia, para o qual não existem grandes bases de dados; o uso de técnicas de aumento de dados (*data augmentation*), que não são perfeitas, apresentam um obstáculo ainda maior. Outro problema encontrado é que os resultados do estado da arte raramente são reprodutíveis porque os métodos usam conjuntos de dados proprietários, redes pré-treinadas e/ou inicializações de peso de outras redes maiores. Um exemplo de um cenário de poucos recursos existe mesmo no quinto maior país do mundo em território; lar da maioria dos falantes da sétima língua mais falada do planeta. O Brasil é o líder na economia latino-americana e, como um país do BRIC, deseja se tornar um participante cada vez mais forte no mercado global. Ainda assim, a proficiência em inglês é baixa, mesmo para profissionais em empresas e universidades. Baixa inteligibilidade e forte pronúncia podem prejudicar a credibilidade profissional. É aceito na literatura para ensino de línguas estrangeiras que é importante que os alunos adultos sejam informados de seus erros, conforme descrito pela "Noticing Theory", que explica que um aluno é mais bem sucedido quando ele é capaz de aprender com seus próprios erros. Um objetivo essencial desta tese é classificar os fonemas do modelo acústico, que é necessário para identificar automaticamente e adequadamente os erros de fonemas. Uma crença comum na comunidade é que o aprendizado profundo requer grandes conjuntos de dados para ser efetivo. Isso acontece porque os métodos de força bruta criam um espaço de hipóteses altamente complexo que requer redes grandes e complexas que, por sua vez, exigem uma grande quantidade de amostras de dados para gerar boas redes. Além disso, as funções de perda usadas no aprendizado neural não fornecem garantias estatísticas de aprendizado e apenas garantem que a rede possa memorizar bem o espaço de treinamento. No caso de fala marcada ou com forte pronúncia, em que uma nova amostra pode ter uma grande variação comparada com as amostras de treinamento, a generalização em tais modelos é prejudicada. O principal objetivo desta tese é investigar como generalizações acústicas mais robustas podem ser obtidas, mesmo com poucos dados e/ou dados ruidosos de fala marcada ou com forte pronúncia. A abordagem utilizada nesta tese visa tirar vantagem da *raw feature extraction* fornecida por técnicas de aprendizado profundo e obter garantias de aprendizado para conjuntos de dados pequenos para produzir resultados robustos para a modelagem

acústica, sem a necessidade de big data. Isso foi feito por meio de seleção cuidadosa e inteligente de parâmetros e arquitetura no âmbito da Teoria do Aprendizado Estatístico. Nesta tese, uma arquitetura baseada em Redes Neurais Convolucionais (RNC) definida de forma inteligente, junto com janelas de contexto e uma árvore hierárquica orientada por conhecimento de classificadores que usam Máquinas de Vetores Suporte (*Support Vector Machines - SVMs*) obtém resultados de reconhecimento de fonemas baseados em frames quase no estado da arte sem absolutamente nenhum pré-treinamento ou inicialização de pesos de redes externas. Um objetivo desta tese é produzir arquiteturas transparentes e reprodutíveis com alta precisão em nível de frames, comparável ao estado da arte. Adicionalmente, uma análise de convergência baseada nas garantias de aprendizado da teoria de aprendizagem estatística é realizada para evidenciar a capacidade de generalização do modelo. O modelo possui um erro de 39,7% na classificação baseada em frames e uma taxa de erro de fonemas de 43,5% usando *raw feature extraction* e classificação com SVMs mesmo com poucos dados (menos de 7 horas). Esses resultados são comparáveis aos estudos que usam bem mais de dez vezes essa quantidade de dados. Além da avaliação intrínseca, o modelo também alcança uma precisão de 88% na identificação de epêntese, o erro que é mais difícil para brasileiros falantes de inglês. Este é um ganho relativo de 69% em relação aos valores anteriores da literatura. Os resultados são significativos porque mostram como *raw feature extraction* pode ser aplicada a cenários de poucos dados, ao contrário da crença popular. Os resultados extrínsecos também mostram como essa abordagem pode ser útil em tarefas como o diagnóstico automático de erros. Outra contribuição é a publicação de uma série de recursos livremente disponíveis que anteriormente não existiam, destinados a auxiliar futuras pesquisas na criação de conjuntos de dados.

Palavras-chave: Palavras-chave: reconhecimento de fonemas não nativos, modelagem acústica, aprendizado profundo, Teoria do Aprendizado Estatístico, processamento de fala, visão computacional, redes neurais convolucionais, máquinas de vetores de suporte.

ABSTRACT

SHULBY, C. D. **RAMBLE: robust acoustic modeling for Brazilian learners of English**. 2018. 160 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

The gains made by current deep-learning techniques have often come with the price tag of big data and where that data is not available, a new solution must be found. Such is the case for accented and noisy speech where large databases do not exist and data augmentation techniques, which are less than perfect, present an even larger obstacle. Another problem is that state-of-the-art results are rarely reproducible because they use proprietary datasets, pretrained networks and/or weight initializations from other larger networks. An example of a low resource scenario exists even in the fifth largest land in the world; home to most of the speakers of the seventh most spoken language on earth. Brazil is the leader in the Latin-American economy and as a BRIC country aspires to become an ever-stronger player in the global marketplace. Still, English proficiency is low, even for professionals in businesses and universities. Low intelligibility and strong accents can damage professional credibility. It has been established in the literature for foreign language teaching that it is important that adult learners are made aware of their errors as outlined by the "Noticing Theory", explaining that a learner is more successful when he is able to learn from his own mistakes. An essential objective of this dissertation is to classify phonemes in the acoustic model which is needed to properly identify phonemic errors automatically. A common belief in the community is that deep learning requires large datasets to be effective. This happens because brute force methods create a highly complex hypothesis space which requires large and complex networks which in turn demand a great amount of data samples in order to generate useful networks. Besides that, the loss functions used in neural learning does not provide statistical learning guarantees and only guarantees the network can memorize the training space well. In the case of accented or noisy speech where a new sample can carry a great deal of variation from the training samples, the generalization of such models suffers. The main objective of this dissertation is to investigate how more robust acoustic generalizations can be made, even with little data and noisy accented-speech data. The approach here is to take advantage of raw feature extraction provided by deep learning techniques and instead focus on how learning guarantees can be provided for small datasets to produce robust results for acoustic modeling without the dependency of big data. This has been done by careful and intelligent parameter and architecture selection within the framework of the statistical learning theory. Here, an intelligently defined CNN architecture, together with context windows and a knowledge-driven hierarchical tree of SVM classifiers achieves

nearly state-of-the-art frame-wise phoneme recognition results with absolutely no pretraining or external weight initialization. A goal of this thesis is to produce transparent and reproducible architectures with high frame-level accuracy, comparable to the state of the art. Additionally, a convergence analysis based on the learning guarantees of the statistical learning theory is performed in order to evidence the generalization capacity of the model. The model achieves 39.7% error in framewise classification and a 43.5% phone error rate using deep feature extraction and SVM classification even with little data (less than 7 hours). These results are comparable to studies which use well over ten times that amount of data. Beyond the intrinsic evaluation, the model also achieves an accuracy of 88% in the identification of epenthesis, the error which is most difficult for Brazilian speakers of English This is a 69% relative percentage gain over the previous values in the literature. The results are significant because it shows how deep feature extraction can be applied to little data scenarios, contrary to popular belief. The extrinsic, task-based results also show how this approach could be useful in tasks like automatic error diagnosis. Another contribution is the publication of a number of freely available resources which previously did not exist, meant to aid future researches in dataset creation.

Keywords: non-native phoneme recognition, acoustic modeling, deep learning, statistical learning theory, speech processing, computer vision, convolutional neural networks, support vector machines.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ASR	Automatic Speech Recognition
ATC	Air Traffic Control
BP	Brazilian Portuguese
CAPT	Computer-assisted Pronunciation Training
CD	Context Dependent
CNN	Convolutional Neural Network(s)
DNN	Deep Neural Networks
F1	F-measure
FER	Frame Error Rate
FLA	Foreign Language Acquisition
GAE	General American English
IoT	Internet of Things
KIT	Karlsruhe Institute of Technology
L2	Foreign Language
LM	Language Model
LSTM	Long short-term memory
LVCSR	large-vocabulary continuous speech recognition
MCT	Multi-conditional Training
MFC	mel-frequency cepstrum
NLP	Natural Language Processing
OOV	Out of Vocabulary
PCA	Principal Components Analysis
ReLU	Rectified Linear Units
RNN	recurrent neural network
RNNLM	Recurrent Neural Network Language Modeling
SER	Senone Error Rate
SGD	stochastic gradient descent
SLA	second language acquisition
SLT	Statistical Learning Theory
SotA	State-of-the-art
VAD	Voice Activity Detection

VC	Vapnik-Chervonenkis
WER	Word Error Rate

CONTENTS

CHAPTER

1

INTRODUCTION

1.1 Setting and Motivation: Pronunciation and Brazilian English

While pronunciation is often considered to be a skill of great value to any language learner (GILAKJANI, 2012), it is often the skill that is pushed aside in the current communicative classroom (EGAN, 1999; SAITO; LYSTER, 2012; LYSTER, 2013). Focus on form and mechanics is downplayed more than ever with the increased focus on meaning, ignoring the frequent demand from students to better their pronunciation or "foreign accent" (HINCKS, 2003; CHUN, 2012; THOMSON; DERWING, 2014).

Brazil is far behind the rest of the world, including local competitors in education. Brazil continues to grow, making itself more visible on the world stage. The eighth largest economy in the world (FUND, 2017) spends an exorbitant amount of taxpayer money on education; yielding little success. Many students are left to find their own means of learning, normally through private tutors. Those who plan to go on to higher education, where most of the educational resources are concentrated, learn quickly that English is the lingua franca for scientific work (SWALES, 2004) and many of these students come to the university with great difficulties in that language.

In Brazil, English proficiency is very low and in a country with an ever-growing number of scientific publications in international journals and conferences, English is of great importance. This makes the notion of an online pronunciation tutor attractive and possibly necessary. However, in the speech processing field, there are few resources in general for non-native speakers (RAAB; GRUHN; NOETH, 2007; VU *et al.*, 2014), much less for Brazilians and these problems start with lack of reliable phoneme recognition (WITT, 2012). A large number of resources exist to help Brazilian students with their writing (SCHUSTER; LEVKOWITZ; OLIVEIRA, 2014), however quality pronunciation training

systems for those learners are practically non-existent. The best examples are tools which compare fixed templates to native patterns(MUNRO *et al.*, 2006; DEMENKO; WAGNER; CYLWIK, 2010). Also most didactic material provided by the larger international publishing houses for English pronunciation training that does exist is focused on "at large groups" or at best some of the larger minority groups present in English speaking countries leaving Brazilian learners without instruction tailored to their needs (SILVEIRA *et al.*, 2009; BAUER; ALVES, 2012). There do exist some books for pronunciation targeted at Brazilians, for example, (CRISTÓFARO, 2015), where the learner must interpret the material and self-diagnose.

This situation is tragic since it is common knowledge that L1 transfer plays the single largest role in the acquisition of new speech sounds (FLEGE; MUNRO; MACKAY, 1995; IVERSON *et al.*, 2003), especially on the segmental level.

According to the PISA 2015 study (OECD, 2015), Brazil was ranked 62nd in the world in reading, 65th in Mathematics and 63rd in Science out of the 70 countries evaluated. This puts the country behind a number of third world countries, including several South American MERCOSUL partners like Chile, Argentina, Colombia, Mexico, Costa Rica and Uruguay (in all three areas) and Peru (in Math). Not only is it behind its local competitors but far below the international average.

The English Proficiency Index (EPI) (FIRST, 2017) ranks Brazil 41st out of 80 countries in the world in English proficiency. This rating puts Brazil in the "low proficiency" category and is considered mediocre among its Latin American peers, right in the middle of the pack and neither top three nor bottom three (see the infographic below in Figure 1, available on the EF EPI website).

This data is surprising to many economists who consider Brazil to be a superpower from The BRICS group. Data from the World Economic Outlook, the International Monetary Fund (IMF) (FUND, 2018), currently lists Brazil as the seventh largest economy. Brazil is also included as a member of the BRICS group, which are consolidated as emerging markets. What differentiates Brazil in this group is its commitment to democracy, leadership in innovation and intellectual property, as well as the rise of the middle class. According to Stembridge (2012), more than 40 million Brazilians rose from poverty to the middle class during Lula's government and their ranks grew 64% in a similar time frame and Brazil is 13th in the world rankings in scientific publishing. While these figures are clearly great motivators for Brazilian students looking to become players in the global marketplace, there is another side to that coin which Brazilians know all too well. The country ranks first in the world in expenditures per student at different levels and it's tertiary education institutions enjoy three times their share over primary and secondary education (OECD, 2015) where, as far as English skills are concerned, it is needed most. Due to the poor results which Brazilian students obtain for their tax money, many stu-

Figure 1 – 2017 EPI Rankings of English Proficiency with focus on the region of Latin America

dents are left to "fend for themselves". This has created a large market for private tutoring in Brazil and most Brazilian students who plan to go to college will pay for private tutoring at some point in their early schooling career. For those looking for an affordable and flexible way to supplement their education, the emerging online learning trends including Moodle, Schoology, MOOCs, etc. are certainly attractive and come with a series of advantages for many learners (SHULBY, 2013). It should also be pointed out that online education does not always deliver on its promises and the need for tools that focus on student's speech in foreign language courses is well documented and justified (HINCKS, 2003; CHUN, 2012; WITT, 2012; THOMSON; DERWING, 2014).

There is also an urgent need to develop tools that meet the current needs of Brazilian students who see English as an important skill to have for the advancement of their academic and professional careers (BAUER; ALVES, 2012; CRISTÓFARO, 2015). English is by far the most frequently used language in academic writing worldwide (GENÇ; BADA, 2010). In Brazil, the case is no different. Many of the most prestigious journals and conferences designated with high Qualis ratings, the system for qualification of scientific publications (SOUZA; PAULA, 2002), by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior/ In English: Coordination of Improvement for Higher Level Academics) either include English as an official language or make it the required language. These facts alone are enough to at least extrinsically motivate most students in Brazilian higher education institutions to learn English. However, with regard to the ability to

communicate globally, Brazil is listed at a much more modest position. According to a survey by (PEARSON, 2014), Brazil ranks 71 of the 77 countries studied, with regard to English fluency in a business environment.

Other than these obvious deficits in education it is important to consider the perceptions that such students suffer from. Berger and Calabrese (1975) write that linguistic "first impressions" impact past, present and future alliances. It has been said that it takes seven seconds for humans to judge each other (BLAIR, 2013). A big part of these "judgments" is based on socio-linguistic perceptions (LABOV, 1963). The most noticeable of these is pronunciation (LADEFOGED; JOHNSON, 2014). (RYAN; CARRANZA; MOFFIE, 1977) found that between classifications (nine varieties) of American English with a Spanish accent, those most strongly pronounced were classified as less favorable. A researcher faces this issue when presenting work at international conferences, where even the best researchers will not shine to their full brilliance if they cannot properly articulate themselves in English and the better their English, the more competent they are likely to be perceived as by their colleagues.

Most L2 (second language) pronunciation errors occur due to interference from L1 (FLEGE; MUNRO; MACKAY, 1995; IVERSON *et al.*, 2003). For example, a Brazilian speaker may assimilate the English /θ/[1] or voiceless "th" sound as an /f/ for a word like "think", pronouncing it as "fink" [f i' n k]* or add an /i/ to the end of the word with a final stop which not only resyllabifies it, but ends up transforming it to "thinky" [θ i' n. k i]*. A speaker may even commit both of these errors in the same word to create a word like finky [f i' n. k i]*; thus, reaching a point where the word "think" is not longer actually understandable to a native speaker. It is also clear that input alone is not enough to "notice the gap" (SCHMIDT; FROTA, 1986).

1.2 Gap: Acoustic modeling for accented speech

While most students voice a desire to work on their pronunciation (DERWING, 2003; MUNRO *et al.*, 2006; LANG *et al.*, 2012), the modern trends of the communicative classroom often push these needs aside in favor of intelligibility (LANG *et al.*, 2012; MCCROCKLIN, 2014). However, pronunciation issues can lead to severe problems in communication when engaging speakers who are not accustomed to foreign speech and even great prejudice (RYAN; CARRANZA; MOFFIE, 1977; DERWING, 2003). Currently, there is no automatic solution for Brazilian students learning English as a foreign language, despite the general deficit of English language skills (FIRST, 2017; PEARSON, 2014).

While great advances have been made in Automatic Speech Recognition (ASR) for

[1] Note that all transcriptions in this project utilize the IPA (International Phonetic Alphabet). For more information please refer to: International Phonetic Alphabet Handbook (1999).

native speech (HINTON *et al.*, 2012; LEI *et al.*, 2013), non-native speech has remained a difficult nut to crack (WANG; SCHULTZ; WAIBEL, 2003; VU *et al.*, 2014).

An example of a traditional ASR architecture, taken from Gales, Young *et al.* (2008), can be seen in Figure 2. In this pipeline, the speech signal is the input for feature extraction. Once the feature vectors have been created, they are passed on to the decoder to do the heavy lifting. The models are sequential because the output of one serves as the input for the next. The first model is the acoustic model which matches the features to the acoustic templates available. This is usually done on the frame level (normally 25ms with a step of 10ms) and the model will output classifications and/or probabilities that each frame could be classified as some phoneme. A phoneme is an underlying representation of a phone, the smallest linguistic unit of a speech sound. Phonemes usually correspond to an orthographic letter or cluster of letters like the letter "a" which could be pronounced as the phoneme /a/ or the letters "th" which could be pronounced as the phoneme /θ/. This will be better explained in the following chapters. Once the frames have been classified as probable phonemes, the pronunciation model/dictionary matches possible phoneme sequences to words in the dictionary. Those words are then sent to the language model which matches the probable word sequences to probable sentences and filters nonsense sequences. The output of the language model then gives us the recognized speech in the written orthographic form we are accustomed to see.

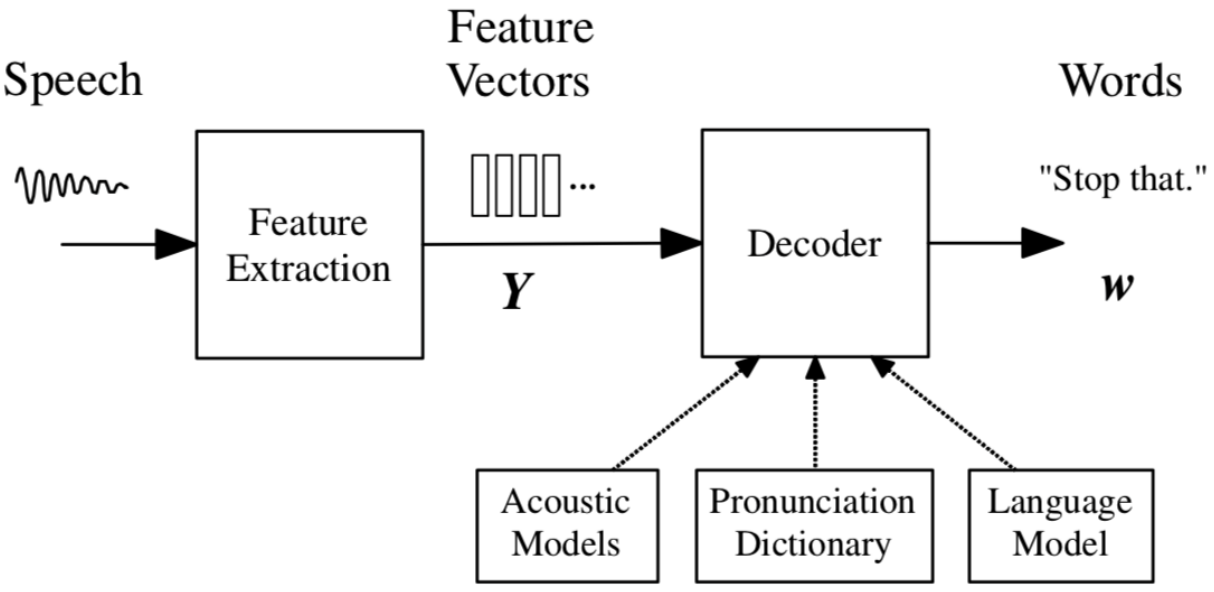

Figure 2 – A Traditional ASR Architecture from Gales and Young (2008)

The difference in Word Error Rate (WER) still seems to be somewhere in the market of 30% in the best scenarios (Hinton et al., 2013 ; Lei, et al., 2013; Wang,Schultz and Waibel, 2003; Vu et al., 2013). This is largely due to the gap in training data between native and non-native speech and the quality of that data. Without large well annotated speech corpora, non-native speech recognition suffers. This can also in part be

attributed to the State-of-the-art (SotA) Acoustic Models (AM) built for native speech using large Deep Neural Networks (DNN) which offer poor generalization (ZHANG *et al.*, 2016; SHULBY *et al.*, 2017) for speech with very different characteristics from the training data. The goal of this thesis is to bridge this gap for more robust phoneme recognition which could be applied to automatic pronunciation training or other tasks were a high accuracy for phoneme recognition is required, including general purpose ASR.

1.3 Objective: Robust acoustic modeling for low-resource scenarios

The objectives of this thesis can be viewed on two levels, one is to innovate and contribute to the state-of-the-art acoustic models, especially on the phoneme level and robustness. The second part of the objective of this thesis is on the applied level where pronunciation tutors need the resources to properly asses the pronunciation of learners and in a special case for the needs of Brazilian students learning English which is used as the raw material for this project.

In order to reach these goals it is important to take research in several areas into account, for example: General American English (GAE) and Brazilian Portuguese (BP) Phonetics and Phonology; Oral Corrective Feedback; Automatic Phoneme (and in a broader sense Speech) recognition; and machine learning. Therefore, this section will be divided into two areas: 1. Foreign Language (L2) Acquisition and Education; and 2. Speech Processing and Machine Learning.

L2 Acquisition and Education

This study builds on the work from Almeida (2016), focusing on approaches to build more robust non-native speech models; therefore, uses the mispronunciation errors presented in that dissertation which were based on the work of Cristófaro (2015). In total, nine types of errors were selected for this study, as they were deemed to have the most coverage based on the phonetic inventories of BP and English (WEINBERGER, 2014) and are known to cause confusion for Brazilian learners. They are:

1. Syllable Simplification: [st] → [ist] - example: “start” → [‘istart];
2. Consonant Substitution: [θ] → [s], [f] or [t] - example: “think” → [‘f i ng k];
3. Non aspiration of stops in the word initial or stressed position: [k^h] → [k] - example: “cup” → [‘k ah p];
4. Devoicing of final obstruents: [z] → [s] - example: “does” → [‘d ah s];
5. Syllable final lateral voicing [l] → [uw] - example: “feel” → [‘f i uw]; ;

6. Lack of syllable final nasal closure + nasalization of preceding vowel - [ih m] → [ihm]; example: “him” → [‘h ihm];
7. Paragoge of voiced velar stop [g]; [g] [ng] → [ngg] - example: “sing” → [‘s I ng g];
8. Vowel assimilation: [ae] → [eh] - example: “bad” → [‘b eh d]; and
9. Intercontinental epenthesis: morpheme [ed] [d] or [t] → [id] or [ed] - example: “danced” → [‘d ae n s ih d].

In this work, we will henceforth consider only the English dialect of GAE. This dialect can be found mainly in the Midwest and is considered the most widely accepted variety of English (LABOV; ASH; BOBERG, 2005) and that which Brazilians are most accustomed to hearing through media outlets and English language instruction in Brazil. The teaching philosophies found in these works are also compatible with the L2 acquisition studies (CRUZ, 2004; DAM, 2012; ELLIS, 1994; SELINKER; LAKSHMANAN, 1992). In light of these concepts, we consider that adult learners will transfer their native phonetic patterns to the target language; thus, creating phonetic constructions resembling an inter-language.

Speech processing and Machine Learning

An acoustic model depends on several resources for reliable representations of speech to be modeled. The following resources (all made available on github[2]) were constructed for this project:

1. A phonetically rich speech corpus containing both native and non-native audio files and their orthographic transcriptions;
2. Interlingual English pronunciation models (phonetically transcribed dictionaries);
3. A grapheme to phoneme (G2P) converter to transcribe unknown words for semi automatic revision;
4. A plugin for Praat to aid in manual revisions of automatic segmentation; and
5. A phonetic balancer script which was used to build the final corpus.

With these resources, an acoustic model capable of modeling both native and non-native speech was constructed. The main objective for the acoustic model is that it is robust enough to handle native and non-native speech with any variety of student-studying environment and advances the state-of-the-art acoustic modeling reported in the literature. Another objective is to open the transparency for acoustic models by making

[2] https://github.com/CShulby

the models available on the project's website and reporting as much detail as possible like PER and FER. In this same line of thought this project presents statistical motivations to prove the convergence of the models in order to support the claims made about robustness.

1.4 Research Questions and Hypothesis

1. Does the model achieve results which are better than the SotA and bring us closer to the results produced by manual alignment?

    a) We assume that a good high end threshold for phoneme recognition is manual alignment, in this case language and pronunciation model errors are not a factor. The current SotA is assumed to be somewhere in between non-native acoustic models and adapted acoustic models. Metrics like Frame Error Rate (FER) and F-measure (F1) are used on the validation sets and are compared to the SotA found in the literature where applicable.

2. Is the model robust?

    a) Robustness is treated both intrinsically and extrinsically. Intrinsically, the model is treated from a Statistical Learning Theory (SLT) perspective where convergence is shown on both the feature extraction level as well as the final classification. This guarantees that the learning is not obtained by chance. Extrinsically, a specific error, epenthesis, will be analyzed to show how the model performs in a hypothetical task like error diagnosis. This task was chosen because it is a particularly difficult construction for Brazilians (KLUGE *et al.*, 2007; MARTINS *et al.*, 2012; JOHN; CARDOSO, 2017) and an error which greatly impedes communication.

The hypothesis of this research is the following:

Given: 1. Statistical learning guarantees; 2. raw feature extraction; 3. careful parameter selection; and 4. knowledge driven classification, a robust SotA acoustic model for accented speech can be built.

Simulating the intelligence of a successful human foreign language tutor in an online environment, a main point of interest in this project is to understand how a computer vision algorithm and expert knowledge driven phoneme classification can contribute to closing the gap for non-native phoneme recognition. Instead of focusing on statistical adaptations, this thesis places more weight on the robustness of each decision made. In this spirit, it is hoped that a powerful resource for pronunciation related tasks can be achieved.

Methods and evaluation

The acoustic model is built based on the theoretical work from Speech Processing, especially as it applies to practices involving Convolutional Neural Network(s) (CNN) (MOHAMED; DAHL; HINTON, 2012; SAINATH *et al.*, 2013; ABDEL-HAMID *et al.*, 2014) with a heavily influence and adaptations from the literature in the area of SLT (VAPNIK, 1998; LUXBURG; SCHÖLKOPF, 2008) which also motivated the use of the HTSVM, inspired by a synthesis of work done is laboratory phonetics (LADEFOGED; DISNER, 2012) as well as SLT (VAPNIK, 2013).

The model is evaluated both intrinsically and extrinsically where: 1. Intrinsic evaluation is carried out by way of FER, PER, F1 on the validation set and Convergence analysis to show that we are able to generalize well; and 2. Extrinsic evaluation for this scenario will be carried out on a specific example taken from well known L2 errors made by Brazilian learners of English and a thorough analysis of that error in practice will be carried out to show how this type of robust recognition could be applied to the task of pronunciation diagnosis.

1.5 Thesis organization

This PhD thesis is organized in the following way. In Chapter 2, the theoretical foundations will be presented for the various fields of study which make up the current work. This will be divided mainly into two main areas 1.) L2 acquisition focusing on Brazilian learners of English; and 2.) Speech Processing, the components needed within that field as well as statistical learning theory as it applies to speech recognition. Then Chapter 3 will build on those foundations and review the literature of related work in phoneme recognition which is further divided into several subsections in an attempt to present a triage of results to better understand the state of the art in this field, being 1. ASR with CNN; 2. HTSVM for phoneme recognition; and 3. Studies which are most impacted by acoustic model results, mainly presenting forced alignment, PER and FER results. Special attention wherever possible will also be paid to non-native speech/phoneme recognition and statistically or knowledge-driven techniques used to approach this problem. Then, in Chapter 4, the materials and methods will be described. This includes the relevant details about the corpus used for this study and explain the procedures for the raw data treatment and the spectrograms generated from it. After that, the feature extraction via CNN and classification done by the HTSVM will be discussed and this will produce the final two chapters. In Chapter 5, the experiments and results will be presented as well as the convergence analysis for all final models. This is where the robustness of the model will be shown and the general aim to be as transparent as possible will be followed. Finally, in Chapter 6, we will have a final discussion to sum up this work followed by the conclusions,

limitations and speculations about future work. All resources and results produced by this project are freely available to the community.

CHAPTER

2

THEORETICAL FOUNDATIONS

This chapter will explore the theoretical foundations for this dissertation and will help the reader better understand the theoretical logic behind the current approach. It will be divided into three sections. Section 2.1 will discuss foreign language acquisition research and where applicable, explain its relevance to Brazilian English language learners; Section 2.2 will introduce the relevant technologies from the area of speech processing and how the components of those technologies can be used when applied to this case. Moreover, Section 2.3 will explain the use of neural networks for speech with details as well as a special type of neural network, the CNN and how it can be used as a feature extractor module. Then, the SVM as a classifier will be presented in Section 2.3.2.1 and finally, since it is important to show the robustness of the proposed model, Section 2.3.2.2, will introduce Statistical Learning Theory.

2.1 L2 Language Acquisition

The first part of this subsection reviews the literature relevant to foreign language acquisition, specifically where it applies to Brazilian students of English. The remainder of this subsection introduces the theoretical background of L2 acquisition research in general and also how that research relates to the current thesis. Section 2.1.1 will introduce the native phonetic inventories of both BP and GAE to better illustrate the "mapping" differences which occur from one language to the other, at least on the phonetic level.

The areas of second language acquisition (SLA) and Foreign Language Acquisition (FLA) are considered sub-areas of applied linguistics, psycholinguistics and foreign language education (Gass, 2013). These two names are distinct but do enjoy a great deal of overlap in the research literature. For the purpose of this project we assume that they are equivalent, or at least similar enough to have equal impact on the scope of this project. The careful reader should be alerted that I will use L2 acquisition, the more popular term

in linguistics, as an umbrella term for both.

To better define these terms, SLA refers to the acquisition process of a second language, more precisely, the native language of a place where the learner lives and that language is different from his or her own native language. FLA refers to the process of acquiring a foreign language, meaning that the learner does not live in the country where the language is spoken but is learning a language different from his or her own native language.

Without going into great detail, the key purpose for the separation of these two groups is the level of exposure and manner of acquisition of a new language. The learning trajectory and the language learning process remains the same or at the very least very similar. This study focuses on foreign language learners; however, much of the research cited in this section may refer to second language learners in cases where similar assumptions could be made about both groups. The author has been very careful to analyze second language research from a foreign language acquisition perspective. It should also be pointed out that the focus of this thesis is on interlanguage provoked by phonological assimilation which, for adult language learners, regardless of where one lives, will exist between one language and an another, in this case Brazilian Portuguese and American English. In light of this explanation it can be henceforth assumed that L2 refers to the acquisition of a foreign language.

Much of the research in this area is devoted to interaction between native (L1) and a non-native (L2) speakers and the interlanguage that occurs while an L2 is being acquired (SELINKER; LAKSHMANAN, 1992; ELLIS, 1994; CRUZ, 2004; DAM, 2012; GASS, 2013). The research is often very clear to make a distinction between younger and older learners, due to the "Critical Period Hypothesis", by Lenneberg (1967), where he posits that there seems to be a certain period or window in which a language must be learned (in a native way - for the vast majority of humans). This hypothesis is difficult to prove because of the lack of data, but with studies like Fromkin *et al.* (1974) and on linguistically deprived children like Genie and Isabelle, it is widely accepted that native language learning does have a certain cutoff which more or less correlates with the age of puberty. Most learners will agree that learning a new language after puberty or as an "adult" is significantly more difficult and tends to become more difficult with increasing age and that the fewer languages one can speak fluently, the more difficult learning a new language can be.

These positions by experts in the field are important for this project, since it is aimed mainly at an adult audience aged 18-30 (university students) or older. It is safe to assume that we are working with adult learners who already have an L1, the majority of which are monolinguals or partially bilingual and their L1 is Brazilian Portuguese. With this focus in mind we will consider only adult language acquisition theories. These

students, however, cannot all be grouped exactly in the same way, because they come from diverse backgrounds. Some students have had a significant amount of contact with the L2 from a young age (bilingual parents, or they may have lived in an English speaking country for an extended period of time), many have been exposed to a great deal of media from English speaking countries (music, movies, etc.) and on the other side many have had little or ineffective English training in their primary or secondary schooling. In short, it is impossible to understand fully, to what extent and at what ages they were exposed to English. What is for certain is that any user can be assumed to be somewhere on the spectrum from novice to near-native or native proficiency and all hypothesis within this universe must be inferred as well as possible.

Also, because of the critical period we can expect the "errors" of these interlanguages to be "mapping-errors" in character. If a learner did not learn a certain phonological construction before this critical period, it is difficult to learn it afterwards. Nonetheless, one can imitate or approximate these structures based on one's knowledge and perception of the L2 and the native L1, where more or less similar constructions may exist. All of this further complicates the hypothesis space from a machine learning perspective. This is further explained by Hammarberg (1997), who describes this issue as a series of inferences and deductions based on the learner's knowledge of the L2 and the knowledge of the L1 and is sensitive to both factors of attention and automatization.

The result of this mapping (which is unlikely to be perfect in all cases) is what is perceived by L1 listeners as an "accent" (HYLTENSTAM; ABRAHAMSSON, 2000; CRISTÓFARO, 2015). Less successful mapping can be considered a "strong accent" and often perceived less-favorably by native speakers or could even be perceived as a negative socio-cultural indication of intelligence, education, social class or professional success and can even result in a low level of sympathy for that person (RYAN; CARRANZA; MOFFIE, 1977; EISENSTEIN, 1983; FUERTES; POTERE; RAMIREZ, 2002; MUNRO *et al.*, 2006). Well directed feedback on pronunciation and explicit training could reduce the "accent" of an L2 learner.

This notion comes from multiple studies on language acquisition; for example Swain (2005) developed the "Output Hypothesis". This is similar to Krashen's input hypothesis (KRASHEN, 1985); however, the author argues that input simply isn't sufficient. The learner must actually produce the new language in order to learn it to mastery level. This can be especially true with pronunciation where articulatory mapping is achieved mainly through drills, repetition and practice. Schmidt and Frota (1986), who studied an adult learner of BP, found that the learner improved much more rapidly when he was regularly tutored in that language than when he was not. They argue that it is important for the learner to "notice the gap" or, in other words, the teacher must point out where he is not pronouncing the word correctly for him to notice that he has made a mistake

and be able to take the steps to correct it properly with the guidance of his instructor. This line of research in the literature is the key motivation for the need for pronunciation training.

This is further explained specifically for pronunciation training by (CELCE-MURCIA; BRINTON; GOODWIN, 2010), who propose that the teaching of pronunciation of L2 must consist of five phases: (i) description and analysis; (ii) discriminatory listening; (iii) production controlled with feedback; (iv) production with guided feedback; (v) production in the context communication with feedback. This framework and others similar to it are widely used by researches and educators in favor of pronunciation training and form-focused instruction in these situations (LYSTER, 2013; LYSTER, 2015). The first two phases relate to the perception of the phenomenon by the student (as would be supported by (SCHMIDT; FROTA, 1986), (SWAIN, 2005). In the framework proposed by the authors, the learning begins in the description and analysis of the phenomenon, when the learner is made aware of the existence of the pronunciation phenomenon in question, they are able to learn about it. The authors then recommend a series of activities, including, listening, imitation, phonetic training, minimal pair drills, etc. and then using these types of exercises in the context of real language so that the students, with the help of their teacher can learn by actually using their new knowledge and thus actively constructing it, as in (VYGOTSKY, 1980).

Since learners often produce words and sounds in the L2, which carry similar patterns to those from their L1, we can observe a variety of phenomena in Brazilian learners of English. These errors are explicitly described in 5.

2.1.1 Native Phonetic Inventories

Before we present specific errors, it is important to review the native phonetic inventories of the two languages. It is worthwhile for the reader to note that the absence or inclusion of a sound in parts of this inventory does not necessarily mean that it is impossible and it should be clear that these inventories are meant to serve the language as a whole and are not meant to cover all dialects or regional variations. Phones are indicated by their manner and place of articulation as well as whether they are voiced or voiceless. Areas in gray are judged to be impossible for a given language. These tables are taken directly from the Speech Accent Archive (WEINBERGER, 2014) and as of 2018 can be found online[1]. We will use this as a guideline to establish that certain sounds are likely to be or not to be present for speakers of an L1 and that they may not map 100% to phones in the L2. Since we believe that the learner is most likely to have the greatest difficulty when learning sounds which are not in their native phonetic inventory. By analyzing Figure 3 and Figure 4, one can make at least two obvious observations: 1.)

[1] <http://accent.gmu.edu/browse.php>

several fricative like sounds (specifically, dental fricatives and postalveolar affricates) are included in English and not in Portuguese; and 2.) English has a larger vowel inventory than Portuguese. These observations led us to conclude that even on the monophone level, several English sounds do not exist in any phonemic context in Portuguese. The reader should also be careful to note that the phonology of a language is even more complex than on the phonetic level presented here. Learners are also likely to have great difficulty even with phones which are present in the native phonetic inventory, but occur in different contexts in the L2 than in which he/she is accustomed in his L1, therefore phonological context[2] also plays a large role.

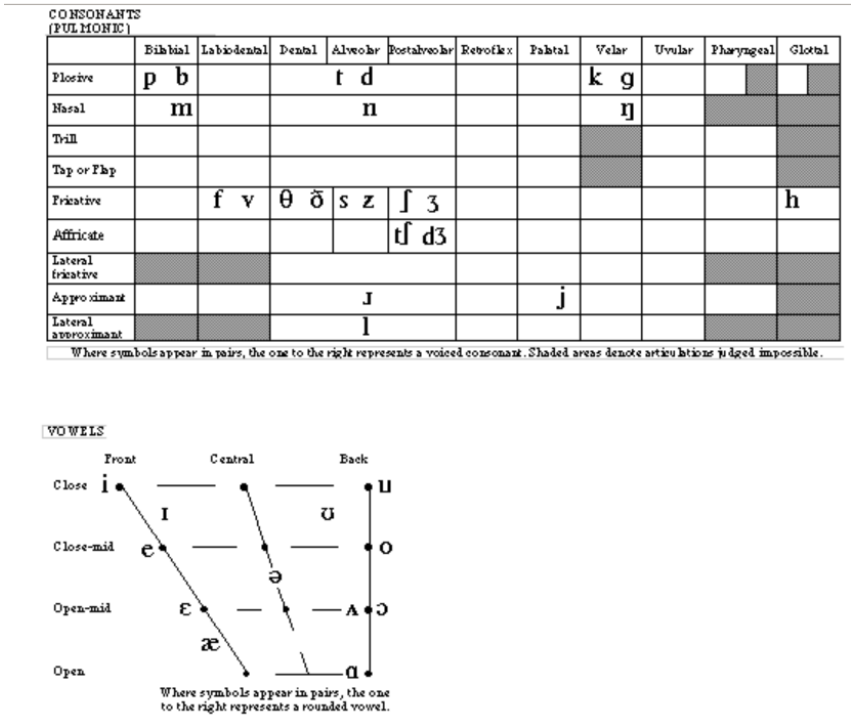

Figure 3 – English Native Phonetic Inventory

[2] Phonological context corresponds to the degree to which a sound segment is appropriate or likely in the context of surrounding speech sounds (MASSARO; COHEN, 1983).

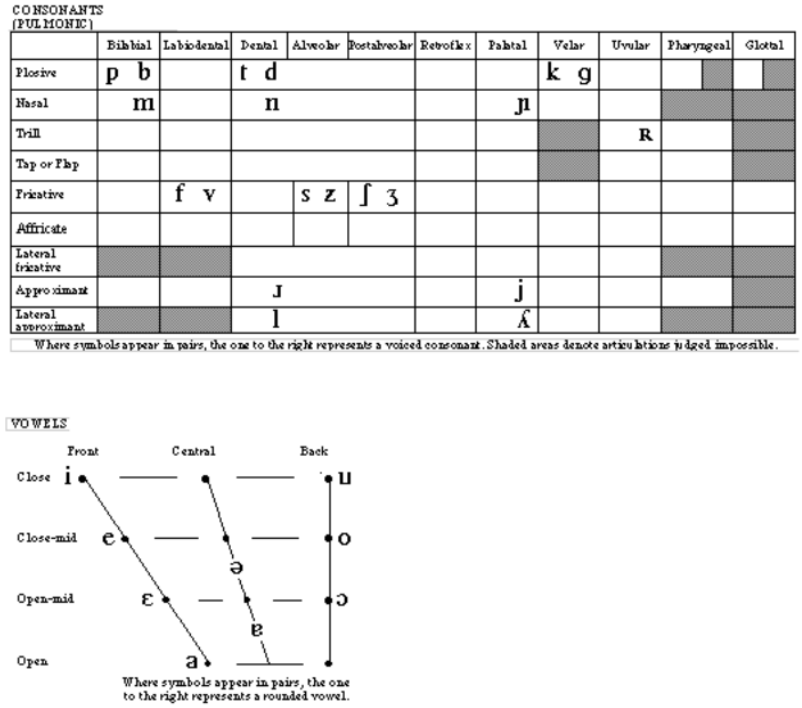

Figure 4 – Portuguese Native Phonetic Inventory

Of course, what becomes more interesting, is how this is manifested at the phonemic level.

2.1.2 Pronunciation Errors

This section will present the errors which could be treated in this scenario. A series of work has been done of Brazilian English pronunciation errors, mainly (CRISTÓFARO, 2015) and has been adapted by Almeida (2016). Figure 5 contains the errors which were adapted from (ALMEIDA, 2016). In particular, attention can be drawn to the first error where syllables are simplified, often with the infamous epenthesis of the /i/ from BP. This example will be expanded and it will be used for extrinsic evaluation. Note that only this error will be explored in this thesis.

Syllable Simplification

Syllable simplification is a phonlogical process which often occurs between morphemes (SATO, 1984; GASS, 2013), some examples are: reduplication, deletion, resyllabification, or cluster reduction (GRAHAM, 2014). When difficult groups of phonemes appear, the learner may try to change the syllable to an easier variety by either adding or

No.	Error	Example	No.	Error	Example
1	Syllable Simplification	[s t] → [i s t] "start" → ['istart] [-k] → [k i] "b ʊ k" → ['b ʊ k i]	6	lack of syllable final nasal closure + nasalization of preceding vowel	[ɪ m] → [ĩ] "him" → ['h ĩ]
2	Consonant Substitution	[θ] → [s], [f] or [t] "think" → ['fiŋk]	7	Paragoge[3] of voiced velar stop [g]	[ŋ] → [ŋg] "sing" → ['s ɪ ŋ g]
3	Non aspiration of stops in the word initial or stressed position	[kH] → [k] "cup" → ['k ʌ p]	8	Vowel assimilation	[æ] → [ɛ] "bad" → ['bɛd]
4	Devoicing of final obstruents[4]	[z] → [s] "does" → ['dʌs]	9	Intercontinental epenthesis[5]: morpheme [ed]	[d] or [t] → [id] or [ed] "danced" → ['dæn.sed]
5	Syllable final lateral voicing	[l] or [ɫ-] → [ʊ] "feel" → ['f i ʊ]			

Figure 5 – Nine Errors - adapted from (ALMEIDA, 2016)

removing sounds to make the word easier to pronounce. For Brazilians this often involves dropping consonants when grouped with other consonants or inserting a vowel between them due to BP's strict VCV (Vowel-Consonant-Vowel) consonant structure. This occurs in the case of start where the foreign cluster [s t] becomes [i s t] rendering a new pronunciation for the word start ['istart].

Epenthesis is also a form of syllable simplification and in Phonology is characterized as the addition of one or more sounds to a word. In BP, this feature is especially marked, where the speaker appends a final /i/ on consonant final words, in most dialects this also triggers palatalization of a preceding /t/ or /d/, e.g. "nerd" > [n eh rx dz i]. Words like ['p i s] 'piece', ['t i m] 'team', ['b uh k] 'book', ['s t a r t] 'start' and ['w er k] 'work' cannot properly be rendered using BP phonological mapping as stops only occur in syllabic onset (the part of the syllable that precedes the nucleus (vowel) of the syllable) and not in the coda (the part of a syllable that follow the nucleus), so that the L2 learner, when dealing with plosives in the final syllable, tends to transfer the characteristics of their L1 to the L2, thereby performing epenthesis and resyllabification in the process. Therefore, these words are often rendered as ['p i: s i] 'piece', ['ch i: m i] 'team', ['b uw. k i] 'book', ['s t a r: ch i] 'start' and ['w er: k i] 'work' For example, Figure 6 shows an autosegmental representation of the word 'book' in standard English pronunciation and pronunciation with the transfer of BP into English.

As noted, the standard pronunciation in the English language the word 'book' is ['b uh k] and the BP transfer variety can be seen here represented as ['b uh: k i]. This may be one of the most difficult sound changes for native English speakers to understand since it is not common for many, non-lusophone, English L2 learners. The syllabification can often render the word completely unrecognizable to the unaccustomed ear.

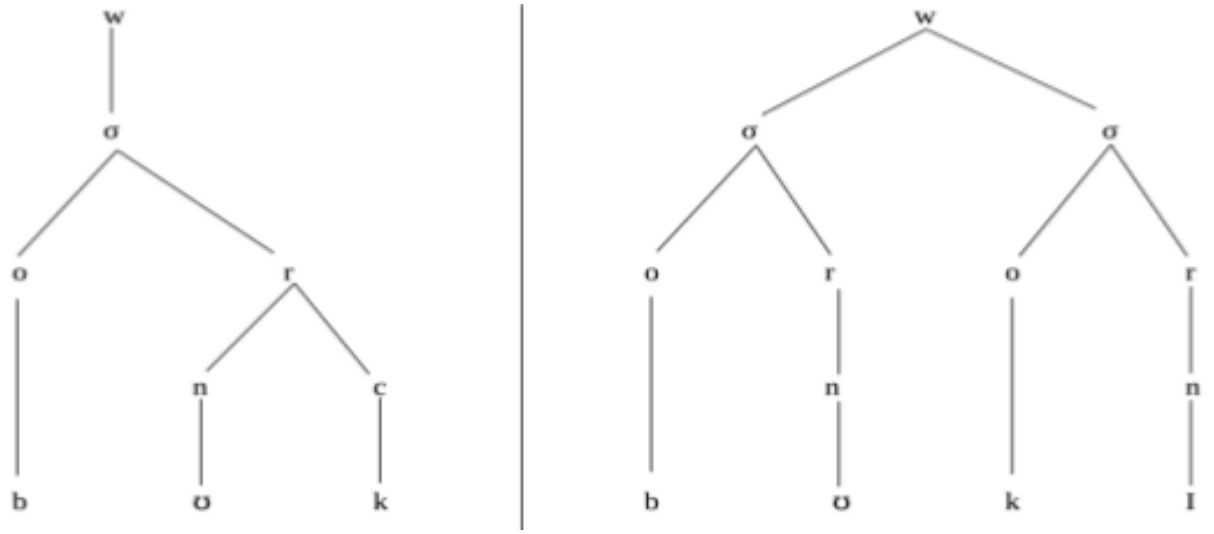

Figure 6 – Realization of the word 'book' in standard English pronunciation [left]; realization of the word "book" with transfer of BP into English [right] - from (ALMEIDA, 2016)

Witt (2012) illustrates some of the challenges for modern Computer-assisted Pronunciation Training (CAPT) systems:

1. Reliable phoneme-level error detection;
2. Distortion error/Accent detection assessment;
3. Text independence;
4. L1 independence;
5. Integrated assessment of both phonemic and prosodic pronunciation components;
6. Corrective audiovisual components;
7. Robust interactive system design.

In this thesis the first item will be the priority. With reliable phoneme-level error detection, the other items may become easier in the future. In our case we are heading in a language dependent direction, but hopefully we will better learn what types of dependencies are really necessary and what is shared.

2.2 Speech Processing

This section is meant to serve as a brief and shallow introduction to the state-of-the-art automatic speech recognition technologies which are related to tasks similar to those contained within this project.

It will focus more specifically on the components which make the most sense for the project's goals and fit within the scope of this thesis. Also an attempt will be made to argue for the theoretical motivation behind the strategies used in this project. The layout of this section is as follows: the next paragraphs will introduce speech recognition in general and explain the linguistic levels of an utterance, specifically the levels which apply to speech recognition and then 2.2.1 will go into further detail about the components required for automatic speech recognition, specifically a.) feature extractors b.) speech corpora; c.) acoustic models d.) language models; and e.) pronunciation models.

Computers are tools which humans use for a variety of tasks and it seems that the possibilities are near limitless. Today we use computers for things we never would have dreamed of twenty or even ten years ago. With the introduction of the Smartphone (Conabree, 2001) and its mass adoption within the last five years we are constantly linked to our devices and this trend does not seem to be slowing down.

ASR is the translation of spoken words into text. In order to understand the speech sounds modeled, it is important to understand how they work in reality. In linguistics, speech is divided into a number of hierarchical levels, each containing the elements of the level below before. These levels have been organized in Table 1 from the smallest to largest units of speech.

Table 1 – Linguistic levels commonly used in ASR

Linguistic Level	**Description**
Phone	An individual speech sound; smallest unit of phonetics For example, in GA "talk" has 3 phones: [t a k]
Phoneme	The basic unit of phonology. An abstract underlying representation of a class of speech sounds which a native speaker identifies as the same sound. For example the rule: /t/ → /tH/ stressed vowel, where /t/ becomes [tH] before a stressed vowel
Allophone	A phonetic variant of a phoneme in a particular language, the basis for narrow phonetic transcription. For example /k/ → /kH/ when followed by a stressed vowel as in "cup"
Morpheme	Smallest unit of a word which can have grammatical meaning. For example the English morpheme [-ed] in the word "hunted"
Word	The smallest unit of a sentence. For example "speech"
Utterance	A communicative discourse unit. For example: "Sally saw Sam".

This is important to understand because smaller units can change depending on the other units near them. For example a phone in one context can appear very different than the same phone in another context. A speaker is influenced by these representations in

his native language, in GAE, when a vowel is followed by an "n", the last few milliseconds of the vowel tends to become nasalized before the consonant /n/ demonstrating a simple case of co-articulation, for example the word: him [h m]. Notice how each phone is still transcribed on its own, even though co-articulation occurs in this phonological context. For Brazilian speakers of English this process is different. Instead of co-articulating the vowel before the consonant, the consonant is dropped entirely and the vowel undergoes a change becoming a nasal vowel, rendering the transcription [h ĩ]. This is something to keep in mind throughout where acoustic models are concerned as this is a difference which needs to be modeled distinctly. Here it also seems appropriate to explain what a "grapheme" is. This term is used throughout this thesis in conjunction with the word "phoneme" where G2P converters are mentioned. A grapheme is the basic unit of orthography. While this term isn't related to phonetics, it is the written form of language which is the output of an ASR engine. With this brief linguistic background in sight, we will now move on to an explanation of ASR systems.

ASR systems range in design mainly with respect to their target audience, usually in reference to the corpus or corpora used in training. Some ASR systems are speaker dependent, which means they were trained with data from one speaker with the advantage of higher accuracy, less training data, but has the trade-off of not having the flexibility to recognize speech from multiple speakers, therefore, decreasing the range of possible implementations. Other systems are speaker independent. These systems require a large variety of well balanced and well-designed data for training in order to be effective. ASR engines depend on data from acoustic and language models, explained below, to perform speech recognition.

In the Introduction, a simple pipeline was shown; in this section I will present more details of each component. It should be noted that we will discuss the traditional ASR pipeline in this section. Even though End-to-End systems have become more popular in recent years, it is uncertain as to how much can be gained by replacing linguistic knowledge with purely statistical models. Furthermore in the current scenario where little data is available, such models do not seem feasible. The typical architecture of an ASR system, taken from Quintanilha (2017), can be seen in Figure 7. Here the raw waveform is the input and the feature extractor extracts the features X from the signal. Those features are then processed in sequential order by the acoustic model, pronunciation model and language model where one can see that the output of one is the input for the next. The problem in most ASR applications is that possibilities can tend towards infinity which means that we have to find a way to maximize these probabilities without needlessly exploding our computational power. Generally this is simplified by the Bayes rules:

$$W* = \underset{W}{\operatorname{argmax}} \frac{P(X|F)P(F|W)P(W)}{P(X)}$$

Here the acoustic model can be see as $P(X|F)$ where it processes the features X and outputs the probable phonemes F which are then the input for the pronunciation model which given the phonemes F, outputs the probable words W. Those words are then sent to the language model which will find the most probable sequence of words as is shown in the output of the system as $W*$.

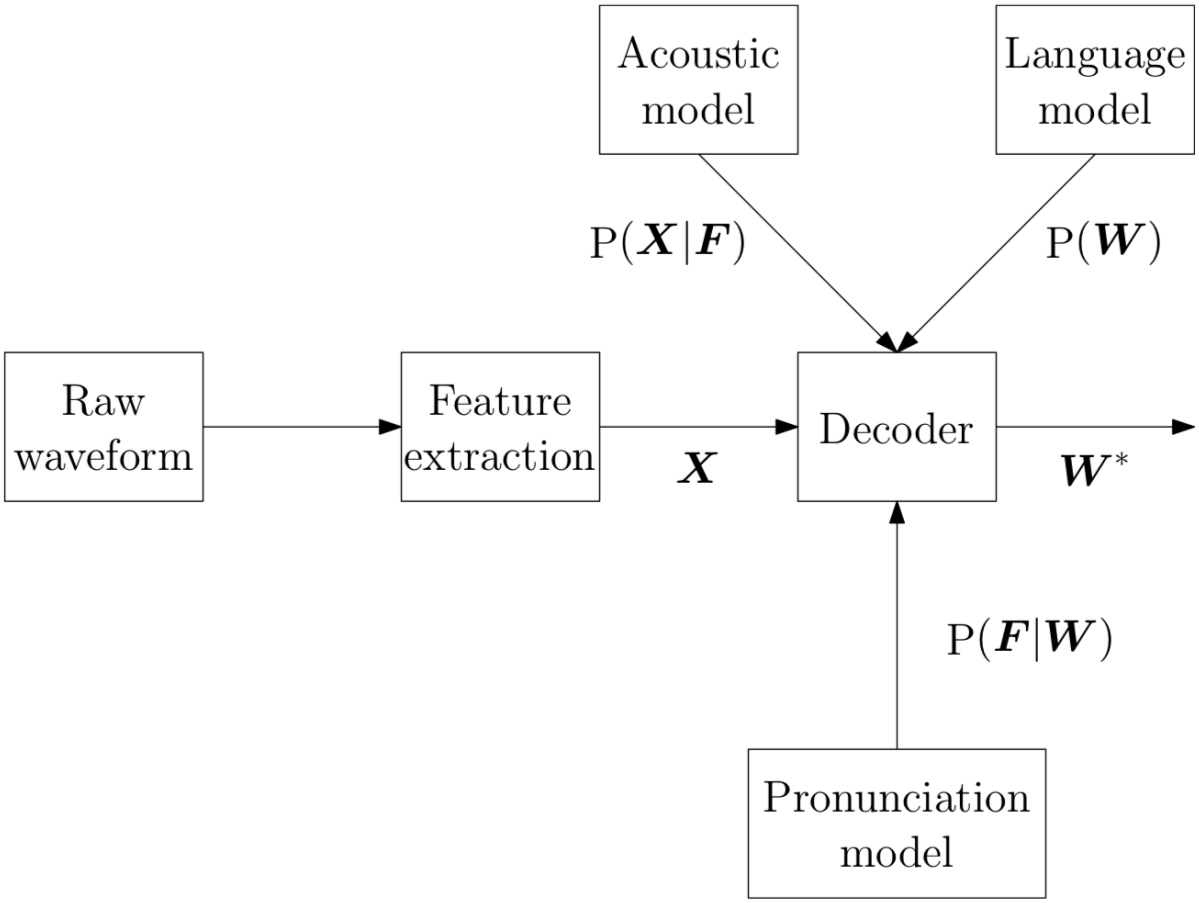

Figure 7 – Typical ASR pipeline from Quintanilha (2017)

2.2.1 ASR Components

This section will introduce the automatic speech recognition components relevant to this project. Not all components may be used in this project, but it is important for the reader to have at least a general understanding of what each of these components are and for what purposes they can be used.

The following components are essential to an ASR system:

1. Speech Corpora
2. Feature Extractors
3. Acoustic Models

4. Language Models
5. Pronunciation Models

Corpora

Non-native speech corpora are the most essential part to an acoustic model; however, they come with two great difficulties for researchers: 1.) finding them and 2.) annotating such corpora adequately. Non-native speech corpora for any purpose are, as imagined, much scarcer than native speech corpora. Even for more popular language pairs, it is difficult to find large corpora which could be suitable for acoustic model training and are often not made available by the usual corpus distributors like ELRA (http://catalog.elra.info/) or LDC (https://www.ldc.upenn.edu/) (RAAB; GRUHN; NOETH, 2007). Here, a number of non-native speech corpora have been identified and a brief overview of what exists in the literature will be provided. This list is not by any means meant to be exhaustively complete, but rather to provide a brief sampling of the most relevant corpora to the current project as well as those which present useful characteristics for phoneme recognition (even if used for different purposes or languages). The following information is provided in the Table 2:

1. Corpus Name to identify the corpus;
2. Target Language (L2) to identify the language(s) spoken in the corpus;
3. Native Language (L1) to identify the influences on the speech samples or "accent";
4. Size to identify the number of hours and/or utterances in the corpus; and
5. Relevant Information to provide any additional information which could be helpful or harmful to phoneme recognition

All corpora come with their own particular design, but not all corpora are created equal. A small change in the purpose of the project could require great changes in corpus design. For example, in the current proposal we are especially looking for:

1. accurate transcriptions on the phonemic level;
2. the best phonemic balancing possible;
3. the largest variety of noises which can be generalized possible.

This creates a bit of a balancing act to find the best corpus possible. What is for certain is that we need to guarantee that we are able to represent the universal hypothesis space or at least approximate as well as possible.

Table 2 – Existing corpora which could be used for non-native acoustic models

Corpus Name	L2	L1	Size	Relevant Information
ATR Gruhn	English	Chinese, German, French, Japanese, Indonesian,	15000 utts.	Rich in variety of speakers, not rich in phonetic sequences needed to build a robust AM.
ILSE	English	German, Italian.	4000 utts. 18 hours	One of the largest NNS corpora available. Phonetically annotated.
ERJ	English	Japanese	68000 utts.	Pronunciation scores.
Hispanic–English	English	Spanish	24 hours	Designed with TESOL Framework. Focuses on proficiency.
Japanese Accented Spanish	Japanese	Spanish	8.6 hours	Various types of speech. Phonetically annotated.
COBAI	English	Brazillian Portuguese	12.5 hours	Only 60% of corpus transcribed. Speech very spontaneous.

The most important factor in the class distribution and the quality of the annotation, but with more classes than a native model it would be fair to assume that in order to obtain the same quality, one would need more data. Still, it seems valid to point out that the numbers of hours/utterances seem to be very small when compared to the corpora currently used for native speakers. For example, the Buckeye Corpus (PITT *et al.*, 2007), a conversational speech corpus, contains high-quality recordings from 40 speakers, totaling 40 hours. It is also orthographically and phonetically transcribed to a precise degree. Recently, larger speech corpora have been released due to interest in deep learning like the Librispeech corpus, an ASR corpus based on public domain audio books (Panayotov et al., 2015). This corpus contains over 1,000 hours of audiobook recordings, although not phonetically annotated. The only known corpus of spoken English by Brazilians is the COBAI corpus (Mello, 2012); however, this corpus doesn't seem to be the best corpus for building acoustic models for phoneme recognition since

1. only about 60% of the corpus is actually annotated;
2. the annotated portions present some transcription issues provoked by the tools used;
3. the speech is broken into 5 minute chunks (very long) and includes many non-phonetic and unannotated related speech issues (long pauses, nonlinguistic sounds, etc.).

In any case, a common assumption and belief would be that one would need more data from non-native speakers to even begin to compare acoustic models. Even then, it is well known that non-native speakers do poorly with native acoustic models and great difficulties have been noted in creating non-native acoustic models (Wang, 2003; Vu et al., 2014). Multiple approaches have been taken to modeling non-native speech including: 1.) native acoustic models; 2.) non-native acoustic models; 3.) bilingual acoustic models; 4.) pooled acoustic models; and 5.) statistically manipulated versions of the prior models.

Native acoustic models are not adequate for non-native speech due to the interference of unknown phonemes from the learner's native language. Purely non-native acoustic models also are considered insufficient for the opposite reason, when one learns new sounds, they are minimally transformed; therefore, unique to non-native phonemes. Wang, et al. (2003) report 49.3% and 43.5% WERs respectively for those types models. Bilingual acoustic models (native speech from both languages) were only slightly better than the native models at 48.7% WER. Of all the corpora types pooled models (native and non-native target language data) were best with a WER of 42.7%. These figures dropped to the mid to low 30's after different statistical treatments which will not be explored in this section as their scope it outside of the definition of the corpus itself. For the interested reader, that work has been continued by Vu et al. (2014) and through statistical adaptation the researchers have been able to obtain at best around 30% WER. It is also worth pointing out that the authors mention that non-native data was sparse and it could be fair to assume that with more non-native data, these baseline numbers could be dropped further.

In light of the studies presented, a list of features for the current corpus can be generated:

1. Native and non-native speech in order to create a pooled corpus;
2. The more data the better;
3. Collection should be taken via readily available equipment to students (e.g., laptop microphones);
4. A variety of language, including more technical information used in academic vocabulary but not solely, a great deal of more basic or intermediate language should be used as well; and
5. Greatest possible phonological variety.

Feature Extractors

An important step in the pipeline is extracting the features from the audio signal. The ASR system needs some kind of statistical representation of this signal in order to op-

erate. Traditionally this is done using mel-cepstral features. The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The result of the MFCC output can be seen in Figure 8

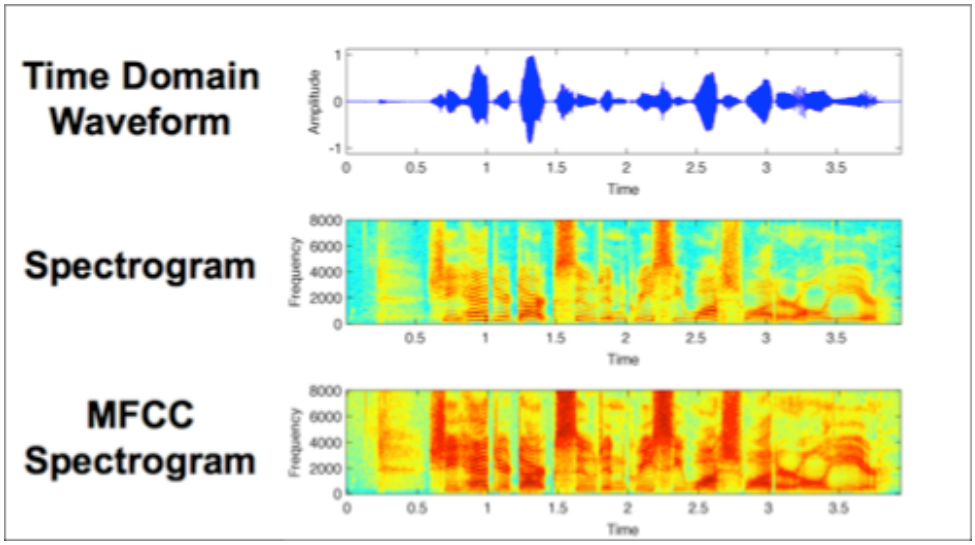

Figure 8 – Comparison of Waveforms Spectrograms and MFCCs from Meza (2018)

MFCCs are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. MFCCs are commonly derived as follows(XU *et al.*, 2004; SAHIDULLAH; SAHA, 2012):

1. Take the Fourier transform of (a windowed excerpt of) a signal;
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows;
3. Take the logs of the powers at each of the mel frequencies;
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal; and
5. The MFCCs are the amplitudes of the resulting spectrum.

These features along with the first order delta and second-order delta–delta regression coefficients plus pitch and the energy of the signal are often used by traditional ASR systems, like HTK (YOUNG *et al.*, 2002) and SotA DNN systems like Kaldi (POVEY *et*

al., 2011). An issue that arises often is exactly how many filter bands to sample. Theoretically, the more the better the resolution but it is also possible that the extra features could contribute to unwanted latency with little advantages. Normally, ASR systems use about 7 to 9 and speech synthesis systems (where high resolution is necessary) will use about 22 to 25. While MFCCs are normally very good at modeling the band frequencies, they do come with some disadvantages. For one, they are notoriously poor at dealing with noise and some researches have tried to find work-arounds to this problem (TYAGI; WELLEKENS, 2005). Also, it is clear that any band-pass approach will be loading quite a bit of non-discriminative information. This is where raw feature extraction can help.

This has also created some interest in the use of full spectrograms and/or waveforms for building feature maps. Recently, Google released Wavenet for speech synthesis (OORD *et al.*, 2016) which uses the waveform as input to a CNN. The output is by far the most natural sounding voice released to date. Also from Google Abdel-Hamid *et al.* (2014) has used the CNN for ASR. This approach seems quite promising since it achieves nearly SotA results and the CNN as a feature extractor should theoretically find the best features to represent the audio signal without feature engineering and are well known to be robust to noise because of its ability to deal with translational invariance and local distortions (LECUN *et al.*, 1998). More about the CNN will be covered later in this chapter.

Acoustic Models

An acoustic model is a statistical representation of the speech signal. Phonemes or other speech chunks are modeled according to their speech features. Normally this is done using Mel Frequency Cepstral Coefficient (MFCC), spectrograms, or other relevant information found in the speech signal where statistical representations of speech sequences are matched to their respective audio signals. The speech corpus is the input data for the AM where the audio recordings of speech and their orthographic transcriptions are used to build statistical representations of the sounds that make up each part of the transcription. Normally, a pronunciation model or manual phonemic transcriptions are provided to intermediate between the speech signal and the orthographic transcription.

The type of statistical representations in the model will depend on the technique and features used to create it. There are many popular toolkits for acoustic models including the HMM based toolkits like HTK (YOUNG *et al.*, 2002) and CMU's Sphinx (SAMUDRAVIJAYA, 2010). More recently, DNN based toolkits have gained popularity, for example: KALDI (POVEY *et al.*, 2011). A typical HMM model is shown in Figure 9. Here we will use it to model a triphone (the co-articulation of three phonemes). Here we can see that there are three probabilistic states, each representing at least part of a phoneme. In an ideal example, if we assume that the middle phoneme is P, then S1 would

represent the last half the phoneme P-1; the phoneme P would be represented entirely in S2; and the first half of the phoneme P+1 would be represented in S3.

This is a good way to model phonological contexts since a phoneme can be slightly different from the same phoneme in a different phonological context due to the co-articulation of sounds inter and intra-word. For the duration of this triphone the model will decide on a frame (usually 25ms with a step of 10ms) by frame basis whether to stay in the current state (or phoneme) or move to the next state.

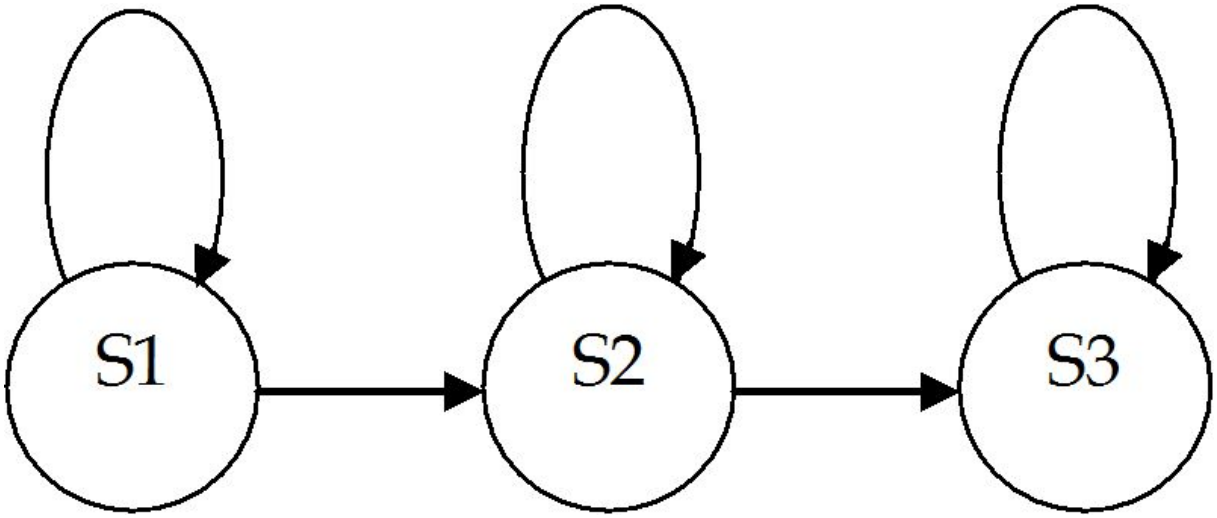

Figure 9 – A typical HMM model

Currently, DNNs are the buzzword in speech recognition. The problem with using a pure DNN structure for acoustic modeling is that Neural Networks are not linear, but speech is. Most systems, including the well-known toolkit Kaldi (Povey, et al. 2011) for example, replace the GMM in the typical HMM based ASR system, creating a hybrid HMM-DNN approach which is now Context Dependent (CD). Figure 10 illustrates this type of architecture well. One can observe that the same HMM triphone modeling above takes place in order to create labels with forced alignment (YUAN; LIBERMAN, 2008). Once the speech is labeled, the DNN will extract the actual features to be used in the model.

This is fundamental for a speech recognition engine as it compares the input signal with the models it has stored in order to make sense of which sounds go with which text representations. Usually, this is done by compiling the recordings (in most cases multiple recordings by different speakers) of a speech corpus and matching each recording with its sentence, both orthographic and phonetic transcriptions are usually given. When trained, the statistical derivations are computed and model templates are stored. Often, a codebook is used to store a finite number of templates with specific features extracted so that the acoustic model contains a limited, but informative and sufficient representation of the speech sounds. Once speech sequence candidates are identified by a system, they

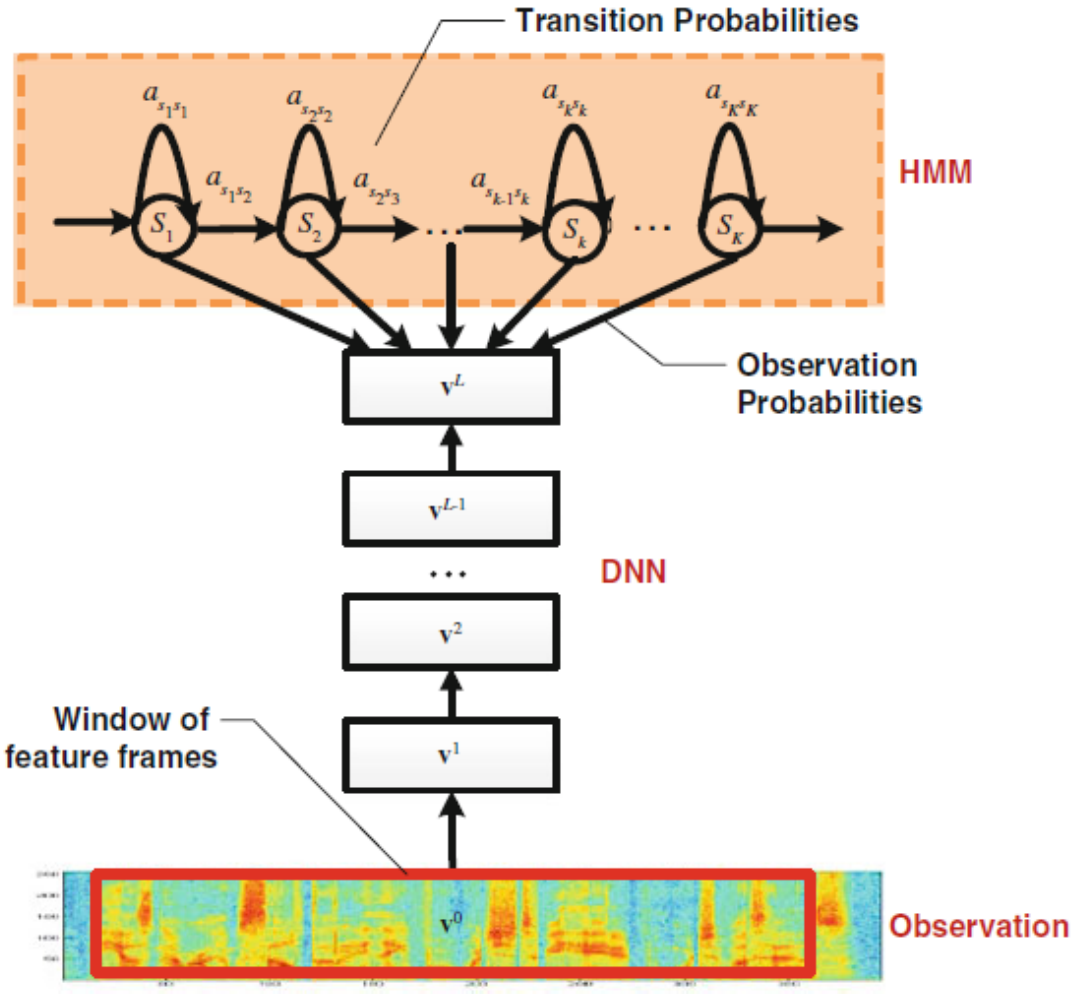

Figure 10 – A typical Hybrid CD-HMM-DNN architecture (DAHL *et al.*, 2012)

are compared to these templates in order to find the transcription which would be the "best" fit.

Language Models

"You shall know a word by the company it keeps" -Firth, J. R. (1957)

A language model is a statistical representation of language (text), or a file containing the probabilities of sequences of words. Since we use a relatively small set of phones which combine into a larger set of phonemes which can create an almost infinite number of sounds. It can become very confusing for the machine to disambiguate the sounds in words, word boundaries and sentences, even in cases, which for humans, could be considered trivial. For spontaneous speech recognition, the accuracy of the recognizer to produce logical sentences is dependent on a language model, defining possible sequences along with statistical representations for their probabilities. This probability is typical computed by the chain rule (adapted from Jurafsky and Martin (2014)) as follows:

$$P(w_1 w_2 ... w_n) = \prod_i P(w_i | w_1 w_2 ... w_{i-1})$$

P("its water is so transparent") =
P(its) × P(water|its) × P(is|its water)
× P(so|its water is)
× P(transparent|its water is so)

There are generally two approaches to building language models that are currently employed in state-of-the-art systems. They are n-gram models, as used in the toolkit SRILM (STOLCKE, 2002) or recurrent neural network (RNN) (see details in Section 2.3) models (MIKOLOV *et al.*, 2010). Both approaches have proven successful.

N-gram models usually require a little more engineering whereas RNN models are more computationally expensive. Still much work needs to be done to improve LM since humans can create an infinite number of "good" utterances with a finite number of words (CHOMSKY, 2014), it is likely that the language model will see many novel but "good" utterances and assign them low scores. To combat this issue techniques like smoothing (CHEN; GOODMAN, 1999) are often employed. Still LM are not perfect and often cause confusion for ASR systems. Complicated grammar structures require a robust language model.

This is where one can see some real problems in the pipeline. The larger the vocabulary is, the larger the number of possible sequences of words and this puts an enormous amount of stress on the LM making it extremely difficult for the system to restrict the possibilities to a manageable number. Academic language can include a much larger quantity of possible grammar structures and many word sequences which would seem unlikely in a web based (chats, twitter) or news based corpora and vice-versa.

Generally this type of problem is dealt with by creating some type of structure to decide which sequences are viable instead of computing all possible sequences. Typically with is done with something like or word lattice which could be compacted further into a confusion network or transducer. Figure 11 shows examples of such a structures.

In order to properly construct an ideal language model for academic English, one requires a very large (millions/billions of sentences) corpus to model in a way that both includes all of the academic areas one wishes to cover and is structured in a way which can easily be understood by a speech recognition engine; in other words, well modeled. Language models are difficult to build for this reason, if strings uttered aren't in the language model, recognition will be low and it is very well known that language from a

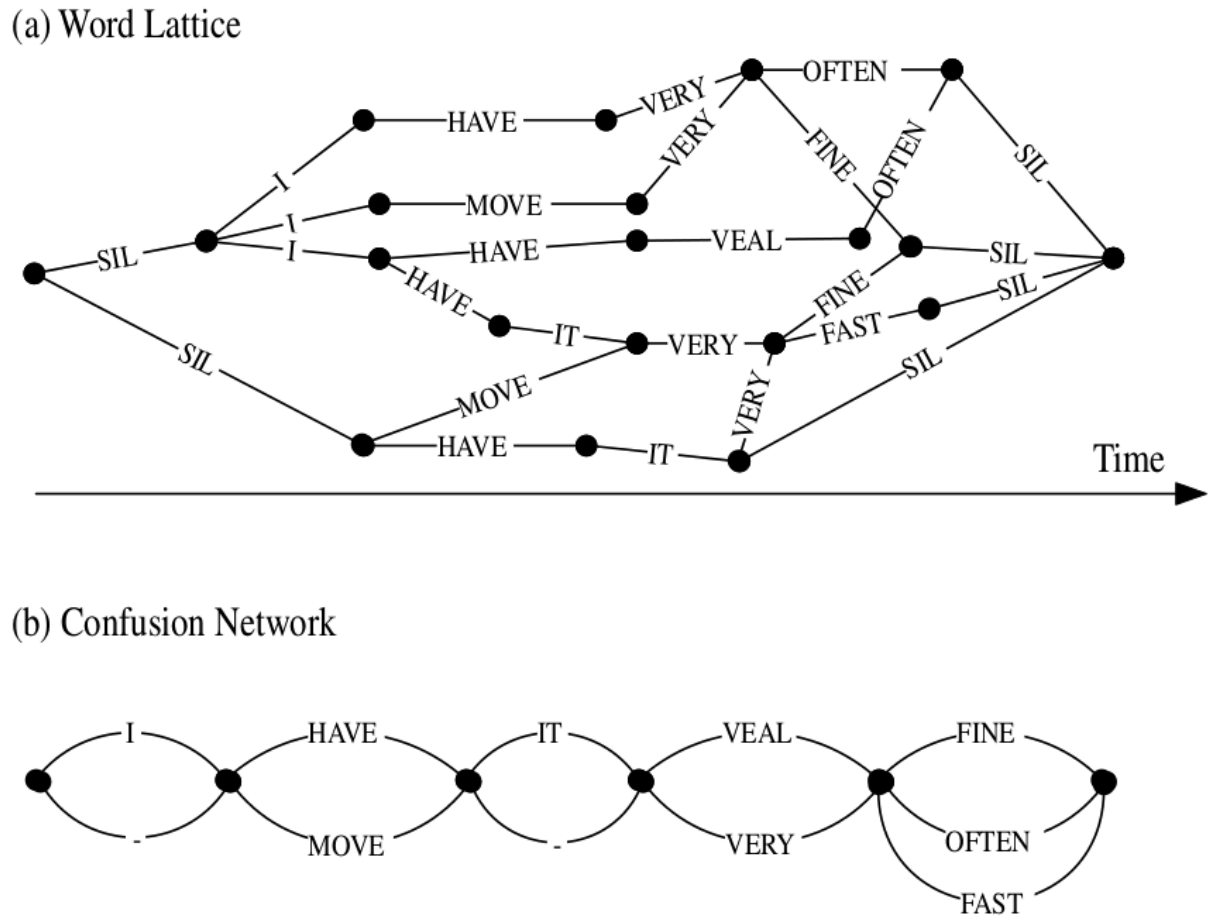

Figure 11 – Typical stricture of a word lattice and confusion network (GALES; YOUNG *et al.*, 2008)

medical journal can be expected to be different than that from an engineering journal, given that the discourse, which the respective communities use, as in different lexical items, in special technical terms or "jargon", tend to be unique (SWALES, 1990).

As if this were not enough to problematize the proposal of creating a language model for this thesis, it should be pointed out that non-native speakers are involved. Non-native language has two principal characteristics that make it difficult to model:

1. Non-native speakers use language (vocabulary and grammar structures) which is typically different from native speakers. Generally, they use structures which are more familiar to them from one or both of two sources: i.) native language, including improbable, possibly impossible word sequences for the target language; or ii.) simplified target language normally more colloquial variations than one would not be expected to use in academic speech. This increases the variability, not only in sheer numbers, but in a way that is difficult to find material and impossible to guess. In the case of Brazilian Speakers, at a minimum, it would require gathering a great volume of work in English written by both Brazilians and native speakers and even then, it is known that spoken language is different from written language,

even in academic settings and even more so when dealing with non-native speakers who would be able to count on native help for proof-reading published works; thus differing from the actual speech they would produce in an impromptu situation. In order to account for this variation, ideally one would need to transcribe large amounts of audio from academic conferences, thesis and dissertation defenses and oral presentations.

2. Variability between non-native speakers is much higher than between native speakers. Academic language includes a series of cultural norms and most native speakers stick very closely to those norms. Non-native speakers have less cultural and language knowledge than native speakers and the range of ability from one speaker to another presents a gamma of variation.

Pronunciation Models

A pronunciation model is the phonetic representation for each word/sentence/utterance analyzed. This can be done with rules, machine learning and/or dictionaries. This type of model is often essential in order to train an acoustic model as it connects the orthographic transcription to its possible phonetic transcriptions, containing words and their possible pronunciations (given as a sequence of phones). The job of the pronunciation model is to connect the acoustic and language model. The language model restricts possible sentences or at least strings of words and the acoustic model possess information about possible phonemes so the pronunciation model connects possible phoneme sequences with possible words which can form possible strings and thus inserted into possible utterances. A typical pronunciation model architecture is pictured in Figure 12. Here one can see the pronunciation lexicon contains words as phonetic transcriptions which are generated by a G2P converter. In this example the G2P converter is a decision tree converter like the one which will be used in this project.

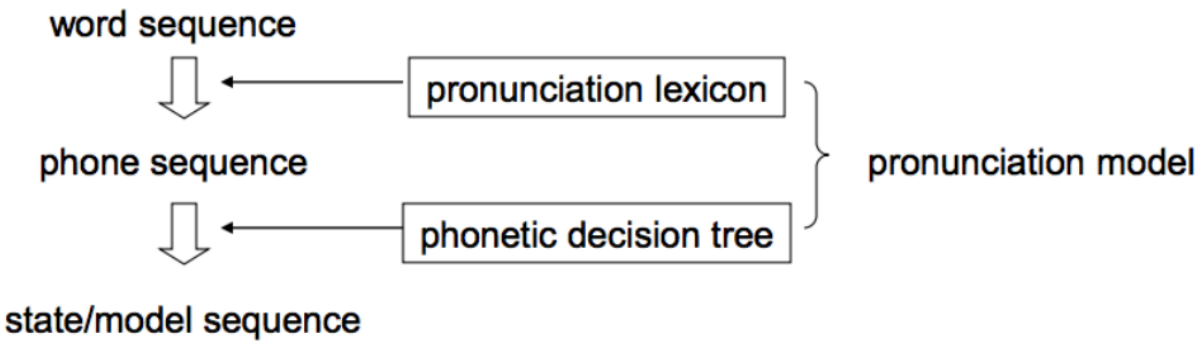

Figure 12 – Typical architecture of a pronunciation model (SCHLIPPE, 2012)

It is important to note that the same word could have multiple pronunciations from dialect to dialect or even speaker to speaker, therefore all possible pronunciations

for each word must be given in these large phonetic dictionaries as options. In fact, it is important that nearly every possible transcription of every word be available in the pronunciation dictionary. This can be done in two ways: 1.) recording all words and saving all corresponding transcriptions in the pronunciation dictionary; and 2.) using rule based expansion of all dictionary entries accounting for all possible pronunciations. It would seem that a hybrid technique would be logical. In an ASR system, the phonemes in the acoustic model are matched with the phonemes in the pronunciation model and only possible sequences are considered as candidates for the recognizer. However, this is not as simple as it may seem as many phoneme sequences can cause confusion over a series of words, for example: "A nice man" and "An ice man". Both of these utterances have exactly the same phonemes. This is further limited by the language model which considers the possible strings of words and assigns scores to the most likely strings. An example of a good, open-source dictionary to use would be the CMU Pronouncing Dictionary[3] (WEIDE, 2005), containing over 134,000 words and their phonetic transcriptions in ARPAbet and is continually being expanded.

2.3 Principal Concepts in Machine Learning

This section is mean to give a very brief overview of the machine learning concepts used in this dissertation. In this section we will go over neural networks, in particular RNN and CNN as feature extractors in Section 2.3.1.1 and Section 2.3.1.2 and then the SVM and SLT which is used for classification and convergence analysis, respectively, in Section 2.3.2.1 and Section 2.3.2.2.

2.3.1 Convolutional networks as feature extractors

2.3.1.1 Recurrent Neural Networks

RNN is a class of artificial neural network where connections between nodes form a directed graph along a sequence. The difference between a RNN and a MLP is that the RNN is able to deal with sequences in an efficient way. With a typical MLP a context window can be given but no additional information about the importance of that sequence is given which requires more complexity to work out. This allows the RNN to exhibit dynamic temporal behavior for a time sequence. Many have termed this behavior as "memory". This type of neural netwok has proven useful in a number of Natural Language Processing (NLP) tasks like handwriting recognition (GRAVES *et al.*, 2009), language modeling (MIKOLOV *et al.*, 2010) and speech recognition(POVEY *et al.*, 2011).

One type of RNN which has enjoyed a great amount of success is the Long short-term memory (LSTM) variation. LSTM networks are most famous for their "forget" gates

[3] More information can be found at http://www.speech.cs.cmu.edu/cgi-bin/cmudict

(GERS; SCHRAUDOLPH; SCHMIDHUBER, 2002). An example of such a gate can be seen in Figure 13. LSTM prevents backpropagated errors from vanishing or exploding (HOCHREITER, 1991). Instead, errors can flow backwards through unlimited numbers of virtual layers unfolded in space. What this means is that the network usually will consider events which happen in close proximity as most important but will still consider long dependencies. This is important for many NLP applications because language often carries dependencies from say a stressed vowel from four or five syllables away (maybe more), a separable prefix as in German, pragmatic context from several sentences earlier or even further.

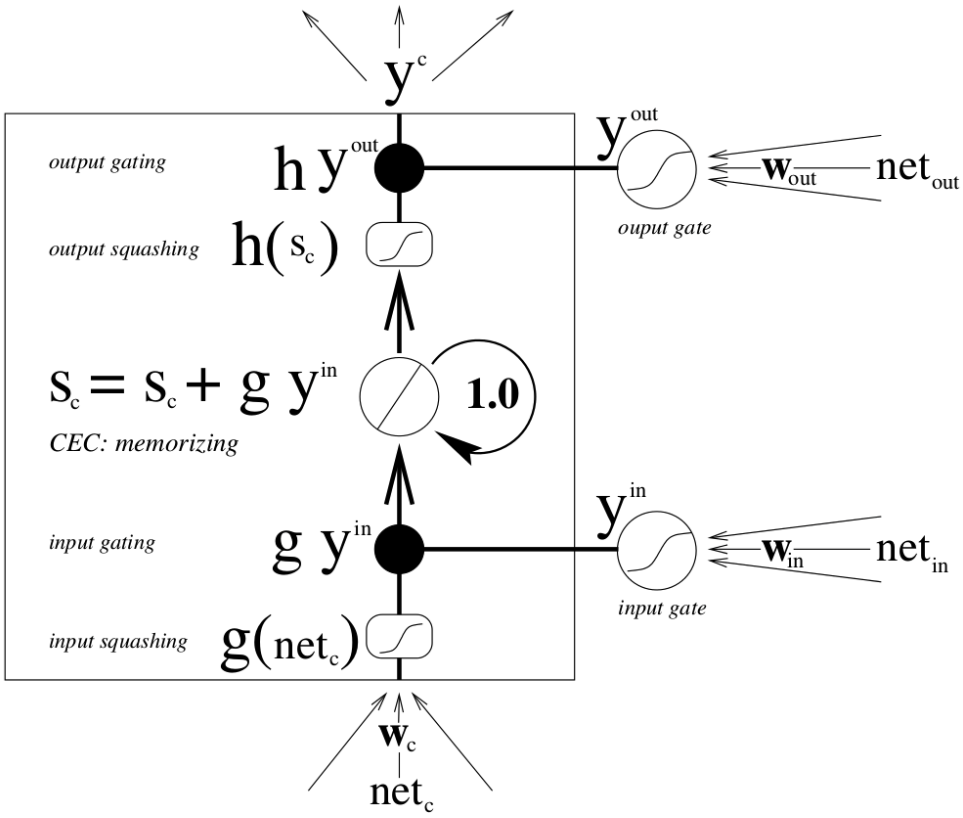

Figure 13 – An example of a forget gate (GERS; SCHRAUDOLPH; SCHMIDHUBER, 2002)

Raw feature extraction through such networks is useful because it is able to regulate which dependencies are worth remembering and which are worth forgetting.

Activation Functions

As an activation function most of the classic works have used the sigmoid function (MITCHELL *et al.*, 1994):

$$f(x) = \frac{1}{1+e^{x-1}}$$

Mitchell calls it the "squashing function", because it is used to compress the outputs of the "neurons" in the MLP and he uses it throughout the book, sometimes referring to it as the logistic function.

Recently Rectified Linear Units (ReLU) have become the new SotA:

$$f(x) = x^{+} = max(0,x)$$

This function avoids negative values and maintains the scale of output values. Basically, every time it finds a negative number, it will be traded out for a 0; mathematically simple and effective. ReLU is also known to train the neural network several times faster without a significant penalty to the accuracy f its inferences (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). It is also desired for the CNN because it increases the nonlinear properties of the decision function and the overall network without affecting the receptive fields of the convolution layer.

2.3.1.2 Convolutional Neural Networks

The CNN algorithm is a biologically-inspired variant of the MLP. Stemming from Hubel and Wiesel (1962) early work on the cat's visual cortex, we know the visual cortex contains a complex arrangement of cells. These cells are sensitive to small sub-regions of the visual field. This concept is called a receptive field and is key in the CNN. The sub-regions are tiled to cover the entire visual field. This is the inspiration for convolutions, a mathematical adaptation to this process. These cells act as local filters over the input space and are well-suited to exploit the strong spatially local correlation present in natural images. This is how the convolutional kernel works in the network. For a more general view of this in practice see the visual in Figure 14. Here it can be seen how a spectrogram from an audio signal is visualized by the CNN where spacial representations are left intact and the kernel convolves over the image to extract pixel information.

The math behind the convolutions is actually quite simple. The trick is having the correct kernel size. A good kernel is one that best represents the avatar to be recognized

in the image. As the kernal convolves over the image it checks for matches, for example: if both pixels have a value of 1 then:

$$1 \times 1 = 1$$

If both have a value of -1, then:

$$-1 \times -1 = 1$$

And if they are different:

$$-1 \times 1 = -1$$

Adding up all of the pixels and dividing them by the number of pixels will generate a filtered map. This is then done repetitively for all of the features and scales linearly with the number of pixels.

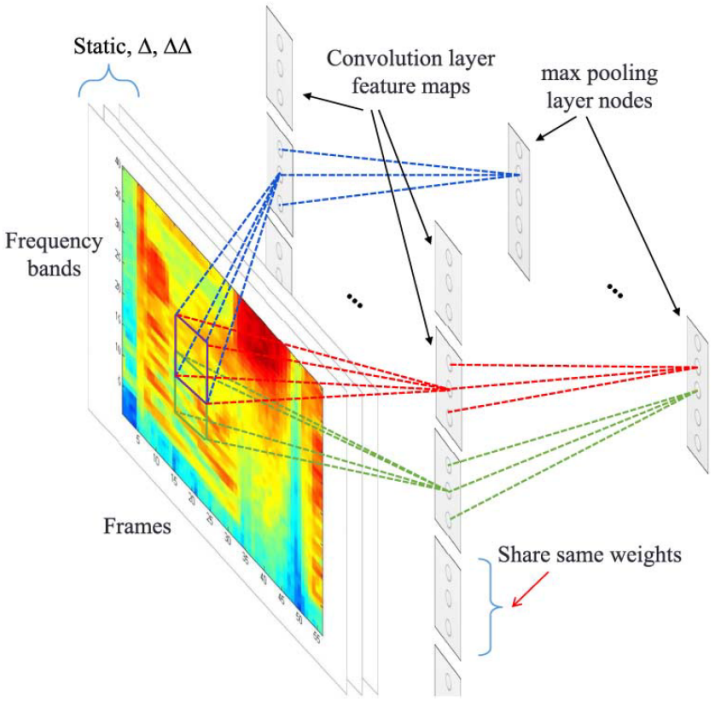

Figure 14 – A typical CNN structure (ABDEL-HAMID *et al.*, 2014)

Additionally, two basic cell types have been identified: Simple cells respond maximally to specific edge-like patterns within their receptive field. This is usually done in the second layer of the network after the pixel information has been processed. This makes sense because mammals are quick to make out general shapes on the presence of objects. Complex cells have larger receptive fields and are locally invariant to the exact position of the pattern. This is like the focus of a mammals visual cortex and after the network has identified the shapes, it will reduce them down to smaller images which are easier to focus on.

Since trained phoneticians are able to do manual segmentation of phonemes, it would seem that a computer vision algorithm would be useful for automatic phoneme feature extraction. In Figure 15, a view from PRAAT (BOERSMA, 2006) as a specialist would see is shown. In this example, taken from the corpus produced by this dissertation, a specialist would revise the boundaries for each phoneme, recognizing the start and finish of each one. This is done by looking at the spectrogram and waveform where information like the explosion of a plosive or the relative formant positions for vowels can be observed and each can be labeled with the correct timestamps.

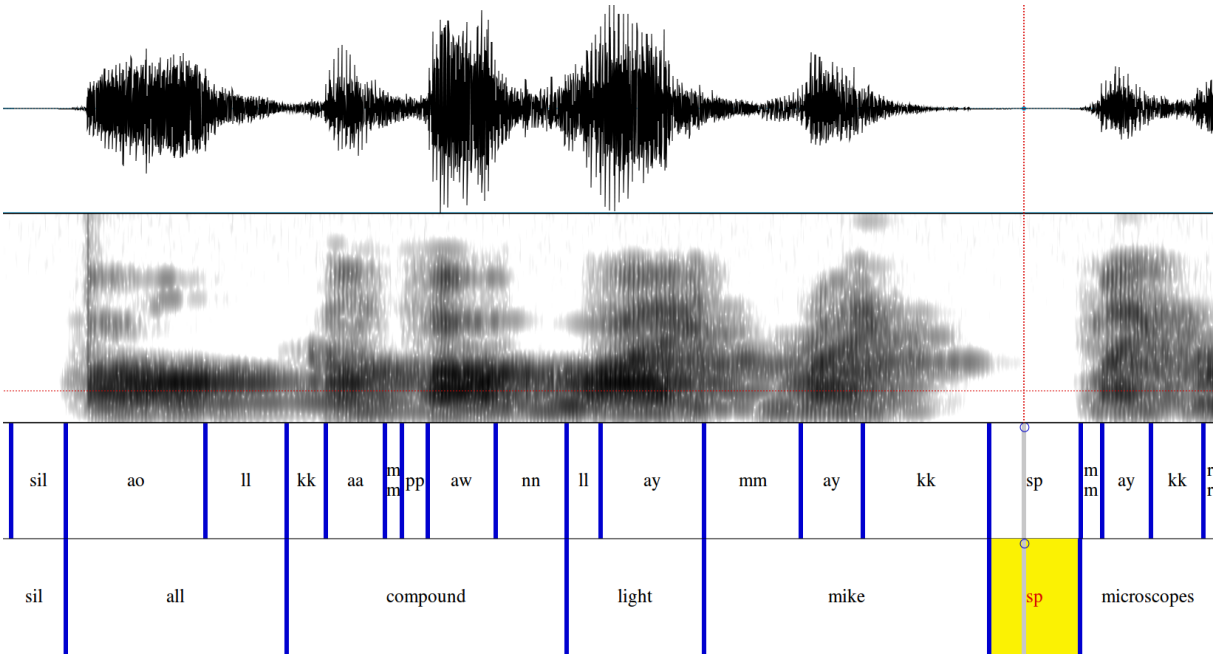

Figure 15 – A view from PRAAT while doing manual segmentation

This is the motivation for the use of the CNN in this dissertation. The use of the CNN was chosen for the following reasons:

1. it is biologically inspired;
2. it has the ability to deal with translational invariance and local distortions;
3. intelligent and raw feature extraction;

4. down sampling of features via max pooling.

Translational invariance

We need to deal with the issue of robustness to noise. Since this is one of the main deficits of the RNN and MFFC feature systems, the CNN provides a great advantage here. The fact that convolutions are able to deal with translational invariance is useful in overcoming this difficulty. The receptive fields preserve the spatial integrity of the image. What this means is that no matter where an object is located in the image or whether it is big or small, light or dark, has been rotated or even partially covered, the CNN can still see that this is the same image as illustrated in the funny example in Figure 16. Jokes aside, this is important because the same logic can be applied to a formant in a spectrogram for example. Depending on the dialect, age or sex of a speaker, formants can occur at quite different frequencies but vowel maps maintain the same relative positions (LADEFOGED; JOHNSON, 2014).

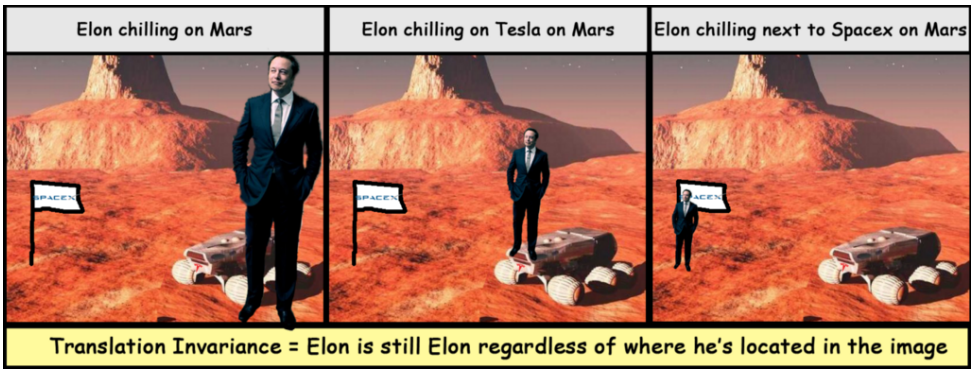

Figure 16 – Elon is always Elon because of translational invariance (taken from Smith (2018))

The spatial size of the output volume can be computed as a function of the input volume size W, the kernel field size of the convolutional layer neurons K, the stride with which they are applied S, and the amount of zero padding P used on the border. The formula for calculating how many neurons "fit" in a given volume is given by:

$$\frac{WK+2P}{S}+1$$

Raw feature extraction

The CNN as a feature extractor has been shown to improve the generalization of robust classifiers like the SVM (WIATOWSKI; BÖLCSKEI, 2017). Since neural networks use raw feature extraction, we can be pleased that this makes our feature engineering job much easier. In traditional GMM-HMM models, one must define a series of features to be used in the clustering of tied states. Also as explained before, the MFCC features are simply band frequency features which will model entire bands even if they won't be used for a certain sound. This is inefficient because we know that certain sounds, like vowels are much richer in information in lower frequencies and fricative information occurs much higher (LADEFOGED; JOHNSON, 2014). Raw feature extraction allows us to focus on the most important features for each class.

Max pooling

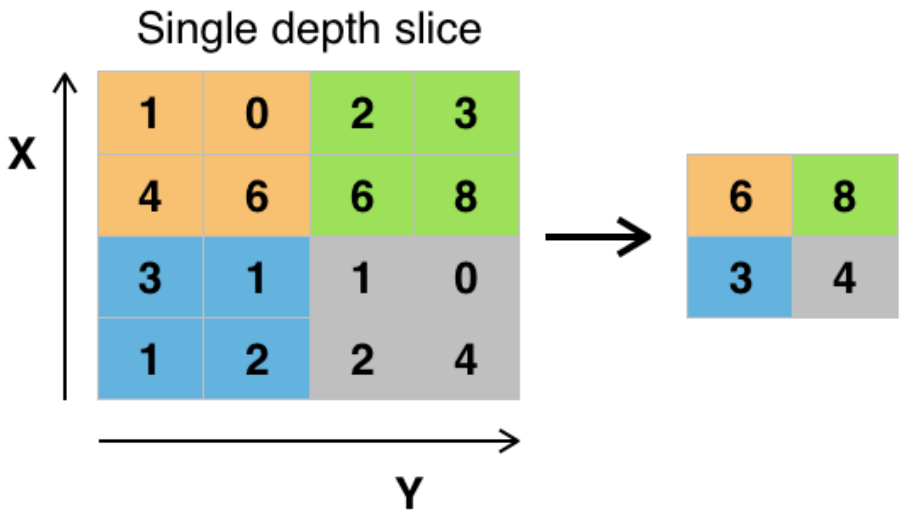

Figure 17 – Max Pooling reduces the feature map

In this case max pooling comes in handy to reduce them to a manageable size. As seen in Figure 17, max pooling is able to reduce features of a determined size along a determined stride. This also greatly reduces the amount of computational time needed which is useful later for classification. A larger pooling size and a larger stride is considered more aggressive as it will output fewer features, whereas a smaller pooling size and smaller strides are more conservative but will output more features than a more aggressive approach. These transformations must be taken into account from layer to layer. It is important to notice that once again the same movement as in the convolutional layer is done to preserve the spatial integrity of the image. As is shown in Figure 17, it moves over the input image as a set of non-overlapping rectangles. The for each of these rectangles, it outputs the maximum. The idea is that the exact location of a feature is less important than its rough location relative to other features. The pooling operation provides another form of translation invariance.

2.3.2 Support Vector Machines and Statistical Learning Theory

2.3.2.1 Support Vector Machines

As a feature extractor, the CNN is very powerful and known to augment the classification potential in the SVM (WIATOWSKI; BÖLCSKEI, 2017), especially due to translational invariance. The problem with neural networks in general is their ability to generalize for unseen sample (ZHANG *et al.*, 2016). This is the strength of the SVM. The SVM is one of the most mathematically robust classifiers in existence (VAPNIK, 2013). It is able to provide supervised learning guarantees from the Vapnik–Chervonenkis (VC) Theory (VAPNIK, 1998) and the principle of structural risk minimization.

In the SVM, we want to train to draw a separation hyperplane:

1. Minimize $d_+ + d_-$, where
2. Distance to closest positive point is d_+
3. Distance to closest negative point is d_-

This can be better visualized in Figure 18. Here one can see the advantage in complexity that the SVM has over a neural network. An MLP, for example, would require at least three hyperplanes to deal with this classification where an SVM can handle it in one. Thus, as long as the features space is adequate, i.e. sufficiently linearly separable, one can employ SVM, which has a more restricted space of admissible functions, while at the same time ensuring learning guarantees.

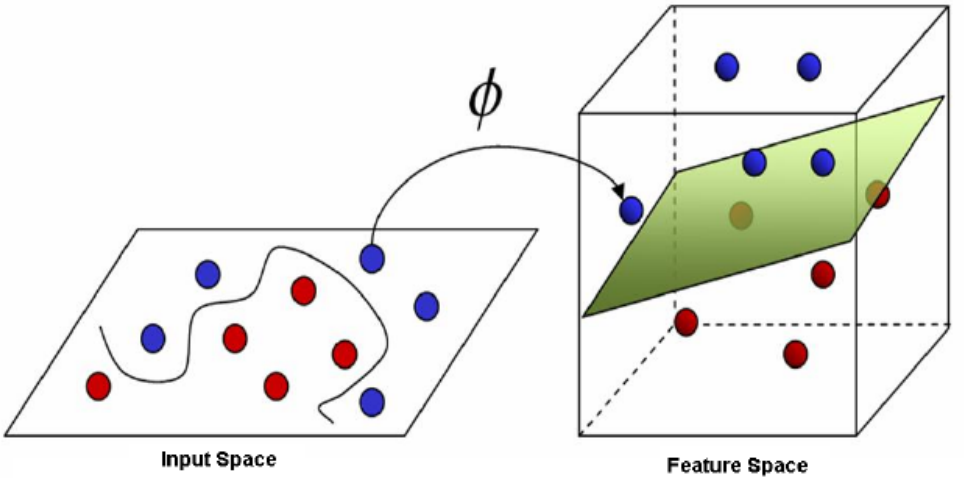

Figure 18 – Kernel transformation from input space to feature space

2.3.2.2 Statistical Learning Theory

Since one of the goals of this project is to provide formal justification for it's robustness in generalization, the decision was made to introduce Statistical learning theory, which provides the theoretical basis for modern day supervised machine learning. The origins of this field go back to the development of the SVM and was popularized by Vapnik (VAPNIK; CHERVONENKIS, 1971; VAPNIK, 1998; LUXBURG; SCHÖLKOPF, 2008). The important fact about this field is that it attempts to draw solid conclusions about the real world from empirical data by approximation. Theoretically, the closer we come to estimating the actual risk, often unavailable, by approximating it via the empirical risk, which can be assessed using the training set, the better we can generalize real world data in the universal hypothesis space.

SLT considers a joint probability distribution over the function P over:

$$P(X \cdot Y)$$

This, of course, can mean different things to different people so to clarify: In Statistics, it corresponds to the relationship between two random variables X and Y, while in Machine Learning, it corresponds to the relationship between input space X and labels Y.

In order to do this, some assumptions need to be made (VAPNIK, 1998):

1. Examples are sampled in an independent manner,
2. No assumption is made about P,
3. Labels can assume nondeterministic values,
4. Distribution P is static, and

5. Distribution P is unknown at training.

Given some function that maps the input space into the labels space, e.g. f: X -> Y, then, a function f is better than another g if and only if:

$$R(f) < R(g)$$

where R(.) represents the risk of choosing function f for the task at hand. Thus, the best classifier f is given by the smaller value for R(f), i.e., the one that presents the lowest risk. Observe that R(f) measures how good f fits the space P(X,Y) even for unseen data. The problem is that we cannot compute the expected risk once we assumed no previous knowledge about the joint probability distribution P. Without the whole universe of examples, it is simply impossible to guarantee this. Since R(f) for a given classifier f can not be computed, one can employ the concept of Empirical risk minimization, where empirical risk is given as:

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{N} l(x_i, y_i, f(x_i))$$

So, while we are getting closer to something useful, it is important to note that a simple loss function as such may provide a memorizer, which is not necessarily an excellent generalizer. The concept of generalization of a classifier f is as follows:

$$|R(f_n) - R_{emp}(f_n)|$$

In this manner, a classifier generalizes well when this difference is small, i.e., the Empirical Risk is close to the Expected Risk. It can also be said that a classifier with good generalization does not necessarily produce a small Empirical Risk nor even a small Expected Risk. This all depends on how we define our goals. And that is where bias comes in. A strong bias assumes a restricted number of classifiers and we can only classify examples within those bounds, a larger bias gives more options but can be difficult to contain. We also have to deal with variance where a tight fight could perform well on training data but poorly on test data. This is called the Bias-Variance dilemma. What we should look for is consistency, or a set of functions, which allow us to study what happens within infinite sample points. This means a supervised learning algorithm should converge to the best classifier, as the sample size increases. So Vapnik (1998) relied on these to prove that a classification algorithm starts with some classifier and tends to find the best one inside a subspace, as the sample size increases, so that the following two requirements are satisfied:

1. Find a way to ensure that the ***Remp*** is close enough to the ***R***

2. A supervised learning algorithm should converge to the best classifier inside its bias, as the sample size increases

This can be illustrated by two common concepts in Machine Learning, namely:

1. Underfitting – If subspace F is small, then the estimation error is small, but the approximation error is large

2. Overfitting – If subspace F is large, then the estimation error is large, but the approximation error is small

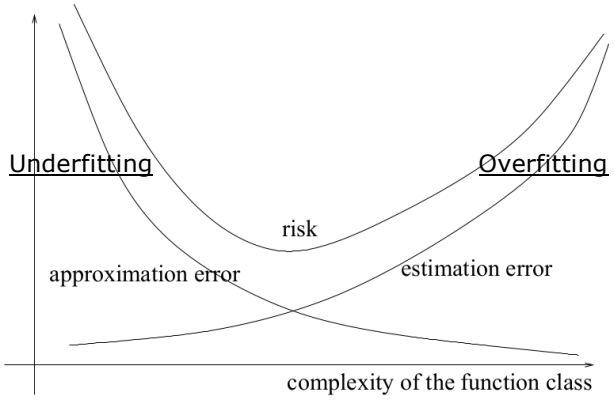

Figure 19 – Figure from (LUXBURG; SCHÖLKOPF, 2008)

The best for supervised learning is the balance. The balance consists in finding a classifier f that minimizes the Expected Risk. However, there is no way to compute R(f) because we do not know the joint probability distribution P. Therefore, we could estimate the Expected Risk using the Empirical Risk and minimize it to find f. To do this we need some kind of a confidence level. In Equation 2.1, the main principle of SLT if defined, which is the Empirical Risk Minimization (LUXBURG; SCHÖLKOPF, 2008). That formulation intends to bound the divergence ε between the empirical risk R_{emp}, i.e., the error measured in a sample, and the expected risk $R(f)$, i.e., the expected error while assessing the joint probability distribution of examples and their respective classes, as the sample size n tends towards infinity. Still, describing the equation, the right-most term is known as the Chernoff bound, f is a given classifier, and $\mathscr{F}$ is the space of admissible functions provided by some supervised algorithm, a.k.a. the algorithm bias (VAPNIK, 2013; LUXBURG; SCHÖLKOPF, 2008; Fernandes de Mello; Dais Ferreira; Antonelli Ponti, 2017).

$$P\left(\sup_{f\in\mathscr{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq 2e^{-n\varepsilon^2/4} \tag{2.1}$$

Vapnik (2013) proved a bound for supervised learning algorithms considering the shattering coefficient $\mathcal{N}(\mathcal{F}, 2n)$, as defined in equation 2.2. Such a coefficient is a measure function to compute the complexity of the algorithm bias, i.e., the cardinality of functions contained in the space $\mathcal{F}$ that produce different classification outputs, provided a sample size n. The generalization bound defined in Equation 2.4, a further result obtained from Equation 2.2 is employed to ensure that the expected risk is bounded by the empirical risk plus an additional term associated to the shattering coefficient and some probability δ (Equation 2.3). This is also known as Vapnik-Chervonenkis (VC) confidence.

$$P\left(\sup_{f\in\mathcal{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq 2\mathcal{N}(\mathcal{F},2n)e^{-n\varepsilon^2/4} \tag{2.2}$$

$$\begin{gathered}P\left(\sup_{f\in\mathcal{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq\delta \\ \delta=2\mathcal{N}(\mathcal{F},2n)e^{-n\varepsilon^2/4}\end{gathered} \tag{2.3}$$

$$R(f)\leq R_{emp}(f)+\sqrt{4/n(\log(2\mathcal{N}(\mathcal{F},n))-\log(\delta))} \tag{2.4}$$

Ideally, We'd like to find a machine learning algorithm able to zero the empirical risk (sufficient capacity) and minimize the VC dimension (capacity not wastefully large).

Note that the constant found by Vapnik is related to both the size of the sample (training set), and the shattering coefficient of the classifier, which is related to the bias of the classification algorithm and can be approximated by its VC dimension. Therefore, a sufficiently large training set and a learning algorithm that does not exceed the necessary complexity for a given problem are the pillars for achieving guarantees in terms of SLT.

CHAPTER

3

LITERATURE REVIEW

In the Theoretical Foundations chapter, the theoretical foundations relevant to this thesis were explored. Since this project's main goal is rooted in robust phoneme recognition using a CNN-HTSVM architecture, this chapter will review the literature of related approaches. Moreover, we will bring forth the most relevant works found in the literature to date and provide a background for this study in order to prepare the reader to understand the contributions made in this project. One of the difficulties in writing this literature review is that while many of the projects components have been used for speech and phoneme recognition, they have not been used together, therefore I try to best illustrate each of these components within the architectures in which they are found in those studies. Another difficulty is that the area of ASR has changed a lot in the last six years. Since 2012, ASR has found a new maturity as deep learning models have made it into the mainstream with everything from virtual assistants like Cortana, Bixby, Alexa, Siri and Google Assistant to smart homes like Echo and Google Home. This has had some side effects for research because most of the studies in recent years have been fueled by the industry. It is common that the principal papers in signal processing conferences are either written by authors already working in the industry (especially Google, Amazon, Facebook, Apple, Microsoft, Samsung... to name a few), or their research and departments are heavily funded by those interests. While this has been very positive for the growth of the area, it also shades a certain translucency over the results presented which make it hard to understand the complete algorithms used, leaving the experienced researcher to do a certain amount of guesswork.

As far as literature review methodology is concerned, I began using the systematic review system but due to the indirect relations of the principal studies on acoustic modeling, a larger focus was placed on the main conferences in this area, mainly INTERSPEECH and ICASSP, along with well known conferences with ACL (Association for Computational Linguistics) proceedings. From there the SotA was established.

A general issue, for proper comparison of acoustic models, is that state-of-the-art methods which use large deep networks with thousands of units and often thousands of hours of training data, do not show frame-level results as in (MOHAMED; DAHL; HINTON, 2012; ABDEL-HAMID *et al.*, 2012; SAINATH *et al.*, 2013; ABDEL-HAMID *et al.*, 2014; GRAVES; JAITLY, 2014; HANNUN *et al.*, 2014; TÓTH, 2015). As outlined in (HINTON *et al.*, 2012), they often employ a number of resources like pronunciation models, language models and other post-processing/data smoothing techniques which are of great help for the end speech-recognition applications; however, they also mask the true recognition accuracy achieved by the acoustic model.

For the reasons stated above, this chapter will be divided into multiple sections and subsections in order to best isolate each aspect under analysis. In the Section 3.1, I will start with the State of the Art in ASR. The Section 3.2 will explore the State of the Art for NN speech. Then the specific components of this project, ASR with CNN and ASR with HTSVM will be explored in Sections 3.3 and 3.4, respectively. Finally, I will try to give some insight about the state-of-the-art research involving acoustic modeling in Section 3.5. Since it is difficult to find work which isolates this module, I will divide it into three further subsections, exploring studies which present results on forced alignment (Section 3.5.1), phone error rate (Section 3.5.2) and frame error rate (Section 3.5.3).

3.1 State of the Art in Speech Recognition

Phoneme recognition is not a new task as explored in multiple studies over the past couple of decades (WAIBEL *et al.*, 1989; LEE *et al.*, 2009; HAU; CHEN, 2011), but greater success has been achieved only in the last five years (HINTON *et al.*, 2012) and still remains far from a solved problem. On the other hand, as previously mentioned, ASR has matured greatly in the last few years. While Phoneme Recognition is a subtask of Speech Recognition and in most systems considered essential, we have to try and separate these two tasks. In Phoneme recognition I want to identify the highest number of phonemes in sequence correctly. When the focus is on automatic segmentation, I may also be interested in identifying the correct boundaries between phonemes and reducing the margin of error as much as possible. Automatic segmentation, often followed by manual revision, is important to most business with ASR products; however, it is not the final task. ASR is the task of converting sound into words automatically. It is important to point out that segmentation errors and phoneme recognition errors are irrelevant if the word is recognized correctly in an end-to-end system. Normally, these types of ”tolerable” errors are corrected by the pronunciation model and language model; normally, they are further refined by post processing rules which can be specifically programmed to attend a certain domain. Logically, most of these studies present WER and few present PER. WER is the most common metric in the literature which makes sense, since the focus is

usually on end-to-end ASR systems and while acoustic modeling improvements are often present it is not the focus for the end result. Studies presenting PER will be isolated later in 3.5.2. As the following subsections go through various angles on the Sate of the Art, it seems useful to start with a table here with each SotA experiment for easy comparison as more detail is presented, Table 3 can be consulted.

Table 3 – SOTA Results for ASR

Task	**Best Study**	**Method**	**Size**	**WER**	**PER**	**FER**	**FA≤10ms**
Google VS	Chiu,et al, 2017	GMM-DNN	12.5k h	5.6	-	-	-
Bing VS	Huang,et al. 2014	RNNLM	24 h	23.2	-	-	-
NIST 2000 Challenge	Soltau,et al. 2014	MLP/CNN +I-Vec	2.3k h	10.4	-	-	-
Broadcast News ASR	Sainath,et al. 2013	CNN	50 h	15	-	-	-
Youtube video transc.	Liao,et al. 2013	DNN +I-Vec	1781 h	40.9	-	-	-
Distant Conversational Speech	Swietojanski, et al., 2013	DNN-HMM	78 h	59.8 w/SM 56 w/MM	-	-	-

Traditionally, HMM (Hidden Markov Models) have been known as the state of the art in speech recognition (MAKHOUL; SCHWARTZ, 1995). These systems can be considered the single greatest advancement in ASR technology because they made large-vocabulary continuous speech recognition (LVCSR) possible in real time and can be trained on any decent desktop machine these days. For decades, the use of HMM trained systems with Gaussian mixture models (GMM) to represent the relationship between HMM states and their acoustic input was practically unbeatable (RABINER; JUANG, 1993; MAKHOUL; SCHWARTZ, 1995; MARTIN; JURAFSKY, 2000; GALES; YOUNG *et al.*, 2008). This was due to their solid mathematical foundations and relatively small computational cost to train models along with a small footprint to store them afterwards; other promising technologies simply required too much computational power to become viable; however, these systems are far from perfect.

For spontaneous speech, HMM-based systems rarely achieve much better than 25% WER (LEI *et al.*, 2013; HINTON *et al.*, 2012) for well trained, clean, English corpora. This means that only three of four words at best are expected to be correct. Recently, there has been a great deal of investment in DNN (Deep Neural Network) varieties for speech recognition (GENT, 2015; HERNANDEZ, 2013; UPADHYAYA, 2013). These models require an enormous amount of training data (300+ hours of continuous speech) and

resources (large scale GPU clusters) (POVEY *et al.*, 2011). Given these conditions, some models had already begun to break the 20%,15% and even around the 10% mark in WER on some tasks (JAITLY *et al.*, 2012; LEI *et al.*, 2013; HINTON *et al.*, 2012) (Lei, et al., 2013; Hinton et al., 2013) This is significantly better but still presents a great deal of errors.

More recently Google has published results on the Google Voice Search (VS) task, using over 12,500 hours of data with a sequence-to-sequence model to bring the WER for that task down to just 5.6% WER (CHIU *et al.*, 2017). This brute force method improved the same task, when the DB only had 5780 hours, using a GMM-DNN hybrid system, where it achieved 11.8% WER (JAITLY *et al.*, 2012). The Bing VS task from Microsoft has taken a different approach, seeking more robustness with CNN (ABDEL-HAMID *et al.*, 2012; ABDEL-HAMID *et al.*, 2014) andRecurrent Neural Network Language Modeling (RNNLM). On that task the results of 23.2% WER (HUANG; ZWEIG; DUMOULIN, 2014) were published, better than the previous done using a CNN at 33.4% WER by Abdel-Hamid *et al.* (2014), using a much smaller scale corpus than Google of 24 hours with high acoustic variability, but larger than the corpus used for the CNN (18h). It is hard to tell from these two studies whether the difference wasn't due to the difference in corpus size as neither explains this process in sufficient detail.

Another task which has become popular due to the rise in mobile phone usage is the e NIST 2000 challenge which uses the Switchboard (GODFREY; HOLLIMAN, 1993) and CallHome (CANAVAN; GRAFF; ZIPPERLEN, 1997) corpora. Switchboard is a corpus of American English spontaneous conversational telephone speech, it is consists of about 2,400 paired telephone conversations between 543 speakers (302 male, 241 female) from various regions in the United States and contains about 300 hours of speech. The CallHome corpus is composed of 120 unscripted telephone conversations between native speakers of English mostly between family members or close friends overseas and contains about 2000 hours of data. The best results on this corpus have been achieved by Sainath *et al.* (2013) and Soltau, Saon and Sainath (2014) with 10.7% WER and 10.4% respectively.

The next task is the Broadcast News Automatic Speech Recognition task which is made up of speech from news content from television and radio transcriptions. The number of speakers in each program and conditions are quite variable. The English Broadcast News Speech Corpus (FISCUS *et al.*, 1997) is often used for this type of challenge. It is a collection of radio and television news broadcasts from ABC, CNN and CSPAN television networks, as well as NPR and PRI radio broadcasts. The corpus consists of 97 hours of data. The best WER on this corpus was 15% done by Sainath *et al.* (2013), using a CNN-based system.

With the popularization of video posted by internet users, in particular, Youtube, another challenging task has become available. For users there are normally two options

to better the accessibility of their audio content in their videos: 1.) Manual transcription; or 2.) automatic video transcription. This is often challenging due to the high variability in the speech itself as well as the quality of the recordings. The best results on this task were reported by Liao, McDermott and Senior (2013) with a 40.9% WER. This was achieved using 1781 hours of data and tested on 6.6 hours. The reported system had 7,000 output states and was merged with a wide hidden layer architecture with 2048 nodes and a low-rank approximation.

Distant conversational speech recognition has become more popular due to IoT (Internet of Things) and is captured using multiple distant microphones. This type of recognition is rather challenging since the speech signals are degraded by overlapping speech, background noise, and reverberation. Some applications include but are not limited to: classroom Lectures, meetings, court hearings, etc. The AMI Meeting corpus (CARLETTA, 2007) has been used for this task. It contains around 100 hours of meeting recordings from three European locations (UK, Netherlands, Switzerland). Each meeting usually has four participants and the meetings are all in English. Many of the meeting participants are non-native English speakers. The training set has about 78 hours of speech and the test and development sets each have around 9 hours. The best results reported on this dataset are 59.8% WER using a single distant microphone and 56.0% WER using multiple distant microphones by Swietojanski, Ghoshal and Renals (2013), using a Hybrid DNN-HMM model.

DNNs have enjoyed an explosion of attention recently and criticism as well (GENT, 2015; HERNANDEZ, 2013; VASILEV, 2015; UPADHYAYA, 2013; ZHANG *et al.*, 2016). It is important to understand these criticisms for a project like the present. The most obvious is the computational power required. DNN research has been near monopolized by major research groups like Microsoft, Google, Amazon, etc. and a few highly advanced University research groups, for example USP's Supercomputer Euler (ALISSON, 2015). They use massive clustering systems, often reporting more than 300 GPUs to perform mini-batch operations (COATES *et al.*, 2013). Anecdotally, I can attest that I have seen centers which use thousands of GPUs to train their ASR systems. This is expensive for smaller research institutions. Aside from the computational cost, they also run high in data costs. Deep learning architectures were always intended for large amounts of data, the more the better. In order to approximate the empirical risk with the expected risk by loss functions, the only way to do so is by memorization. While they have been hailed for their ability to create an end-to-end process with very little feature engineering, this has a downside because, due to raw feature extraction, they learn (memorize) quite a bit from these features and that includes noise. If the system is to be expected to perform independent speech recognition in anything less than a studio vacuum with no noise, it will need not just a lot of data but a lot of variety in the data with all types of conditions possible. Another issue has to do with the results themselves. DNNs are often prone to issues

like overfitting and vanishing or exploding gradients (PASCANU; MIKOLOV; BENGIO, 2012). Also pure DNNs are not temporal therefore some type of feature extraction or pretraining must be performed before hand in order to obtain good results as done in the work by Povey *et al.* (2011). DNN research is often criticized due to the practices of "fine tuning" where various techniques like bottlenecking or dropout are used to achieve better results, often with little hard research behind them. DNNs are often treated as a magical black box which makes them more difficult to understand (VASILEV, 2015).

Moving away from the topic of machine learning approaches, there are a number of things one can do to better the accuracy of a system. Speech recognition is more complicated than recognition of more harmonic sounds (FORSBERG, 2003). It isn't evenly segmented or clean, but rather is full of noise, non speech sounds and full of suprasegmental information. However, one thing that can be said about speech sounds is that not all sounds are created equal. Some sounds are associated with other sounds in more or less specific contexts. Also, not all tasks are equal; depending on the task one wishes to perform the training process can be benefited by certain practices. On the phonetic level, it is known that certain sequences are more likely within a language (ODDEN, 2005; LADEFOGED; DISNER, 2012). If the language has a CVC (consonant - vowel - consonant) structure then a CCC (consonant - consonant - consonant) recognition would certainly seem unlikely. Any speaker knows that some combinations are considered "tongue twisters" because they are known to be difficult to pronounce and depending on the language certain combinations may even be impossible.

Most systems use trigram acoustic models (YOUNG *et al.*, 2002; SAMUDRAVIJAYA, 2010; CERNAK; IMSENG; BOURLARD, 2012) which are generally sufficient, some even use pentagrams (anecdote) which are much more expensive and generally gain little. Some sounds and words are more frequent and syntactic structure rules are generally respected in any utterance. Only certain words can be grouped with certain words because an utterance is expected to have characteristics like subjects, verbs, objects, tense, mode, etc. For this task a language model, as explained in 2.2.1, is often used, granting higher probabilities to more probable sequences in the language statistics. If the language model is well tailored to the task it can be even better. For example, if the task is mobile app management through a restricted series of commands, one can expect a relatively high level of accuracy since very few words and even fewer specific sequences are required and most are recorded in their entirety in the speech corpus. For this task a trigram language model is generally used. The corpus itself should also reflect the speech to be recognized and be sufficiently generic for said purpose, including a great variety of sound combinations to account for a minimal number of each sound combination possible within the context of the task. The more one is able to limit recognition the better. The highest results are obtained when the system is given the most information. For example if the system already knows which words were uttered, it will better recognize the phonemes in

the utterance. Once the system has identified the word it uses a pronunciation model to identify the phoneme sequences possible.

3.2 State of the Art in Speech Recognition for Non-Native Speech

This is far from trivial since the area has become somewhat stagnant is the last few years, it seems that there may be a revival in the coming years with the rise of Internet of Things (IoT) and virtual assistants. We have recently seen that Google Assistant is available for a great number of languages, while their models rely mainly on statistical adaptations, other approaches should follow as Alexa, Siri, Cortana and Bixby gain more languages. I would like to point out that I will not cover adaptive models, despite their popularity, in this chapter at all because they are out of the scope of this project. Most engineers will agree that these models provide an excellent solution when the end user's usage is known but they were never designed to be robust outside the their adapted domains. What I will cover here are the state-of-the-art approaches using large and small non-native and pooled corpora. Here, I will also start with a table containing each NN speech SotA experiment for easy comparison in Table 4 can be consulted.

Tao *et al.* (2016) was able to achieve the best results known for this task using a tandem HMM-DNN, much like the state-of-the-art systems from Povey *et al.* (2011). In this study, they investigate two deep learning architectures, a DNN and a tandem HMM-DNN with bottleneck features, as well as a GMM system and achieve superior performance in ASR over the conventional GMM system. They use an approximately 800-hour large-vocabulary non-native spontaneous English corpus. The best system achieve 23.07% WER. Interestingly, the authors cite their motivation as improvement for CAPT systems but they do not publish their PER or FER; they use a third of their test data for ASR evaluation, a third for automatic grading training and a third for automatic grading testing. The metric used for this was a kappa score between their system and human raters.

For a small corpus of only two hours of training data (JUAN *et al.*, 2015) was able to score a 32.52% WER using a DNN-HMM system with weights initialized using a Restricted Boltzmann Machines that resulted in a deep belief network with 6 stacks. They then fine tuned the system using Stochastic Gradient Descent with per-utterance updates, and learning rate 0.00001. The data was taken from a corpus of 15h of English speech spoken by 24 Malaysians (of Malay, Chinese and Indian origin) collected by some of the same authors at the Universiti Sains Malaysia (TAN; BESACIER; LECOUTEUX, 2014). With many fine adjustments it is difficult to discern whether this approach would be robust for other corpora. The same network achieved 40.70% WER using a purely

Table 4 – SotA Results for NN Speech

Task	**Best Study**	**Method**	**Size**	**WER**	**PER**	**FER**	**FA≤10ms**
Large-scale NN Speech	Tao, et al., 2016	DNN-HMM	800 h	23.07	-	-	-
Small-scale NN Speech	Juan, et al., 2015	DNN-HMM w/RBM init	2 h	32.52	-	-	-
Large-scale Native Speech w/ NN Adapt.	Juan, et al., 2015	DNN-HMM w/RBM init + fine-tuned NN Softmax	2 h	24.89	-	-	-
Bilingual L1-L2 Speech	Vu, et al., 2014	Bottleneck MLP w/IPA Mapping	8.17 h	52.73	-	-	-
Large-scale Bilingual L1-L2 Speech in Domain	Garber, et al., 2017	HMM-GMM	75 h native 20 h NN = 95 h Total	6.76	39.24 SER	-	-

native TED Talks corpus of 118h for training. Both models were tested on the 4 hours of data with the same pronunciation dictionaries (CMU Pronunciation Dictionary with no non-native adaptations). It seems to contradict my experience in this area that a smaller non-native corpus would do better than a large native corpus without overfitting. What is clear is that the only transcriptions provided were from the pronunciation dictionary. This strategy may work to memorize inconsistencies in pronunciation but likely the pronunciation dictionary is being heavily relied upon. The authors also showed a network which was not statistically adapted but was fine tuned with the two hours of NN data on top of the 188h trained native system and achieved 24.89%. This 15% difference is likely due to that fine-tuning.

The last strategy I would like to add is the case of pooling bilingual L1 and L2 data. The researchers from Karlsruhe Institute of Technology (KIT) have been working for many years in this area (WANG; SCHULTZ; WAIBEL, 2003) and while the best results

of the work that has been done at KIT certainly involves statistical adaptations, they are also the state-of-the-art for bilingual models which is also the focus of this project. In Vu *et al.* (2014) they obtained an average WER of 52.73% from Bulgarian, Chinese, German and Indian speakers of English. The system was a front-end bottlenecked MLP seeded with IPA-based phone mappings and was trained on 8.17 hours of data. The corpus was obtained from telephone conversations placed within the USA from 42 male and 21 female speakers. Another well-known study using larger corpora for both pooled and adaptive techniques was done by (GARBER; SINGER; WARD, 2017). In this study, the goal was to improve speech recognition for Air Traffic Control (ATC) systems. Two large databases were used being: 1.) 75 hours of US English audio was taken from the 1997 English Broadcast News Corpus (HUB4); and 2.) 20 hours of German-accented data, which is purely in-domain ATC speech, provided by UFA, Inc., a company specializing in ATC training and simulation. Then, it was tested on 6 hours of German-accented speech, which was taken from the ATCOSIM corpus (HOFBAUER; PETRIK; HERING, 2008), which consists of audio recorded during real-time ATC simulations. Their pooled model, trained with the standard Kaldi recipe (POVEY *et al.*, 2011) achieved a WER of 6.76% which is not surprising, given the large amount of data and the restricted domain. What was more interesting is that the authors did provide Senone Error Rate (SER) results where the same model obtained an SER of 39.24%. The authors do not describe this division well but typically, senones in Kaldi are linked to parts of the triphones, roughly equivalent to phonemes. For our purposes will assume this to be comparable to PER. It is actually common to have an inverse relationship in PER and WER for non-native speech. For native speech PER tends to be lower because the differences in one native phoneme often cause the confusion in WER, whereas for non-native speech the phoneme confusion can be high but typically causes less confusion in the pronunciation and language models as they are able to sort this out by context.

3.3 State of the Art using CNN for Speech Recognition

Since the current project uses a CNN as its core feature extractor, this section will explore the best studies available using CNN, especially for phoneme recognition. Table 5 contains the studies references here.

One of best known studies proving the capabilities of CNN for this task is Abdel-Hamid *et al.* (2012), where a hybrid CNN-HMM model using local filtering and max-pooling in the frequency domain is proposed to deal with the translational invariance problem present in other DNN. The HMM deals with the issue of distortions of over time, while the CNN convolves over the frequency to take advantage of its ability to deal with variation among speakers in this domain. That work is continued in Abdel-Hamid *et al.* (2014). This study serves as a baseline on the TIMIT test set for the state-of-the-art deep

Table 5 – SotA Results for CNN Speech and Phoneme Recognition

Task	**Best Study**	**Method**	**Size**	**WER**	**PER**	**FER**	**FA≤10ms**
TIMIT Phoneme Recognition w/ CNN	Abdel-Hamid, et. al.,2014	HMM-CNN	18 h	34.2	21.6	-	-
TIMIT Phoneme Recognition w/ very Deep CNN	Zhang, et. al.,2014	HMM-CNN	4 h	-	18.2	-	-
NIST 2000 Challenge	Sainath, et al., 2013	CNN	300 h	11.5	-	-	-
Broadcast News ASR	Sainath, et al., 2013	CNN	50 h	15	-	-	-

CNN with a 21.6% PER (Phone Error Rate) where they used the 18-hour Bing voice search data for training and then fine-tuned the network with the TIMIT development set.

In (SAINATH *et al.*, 2013), the optimal CNN architecture is explored including the number of convolutional layers and hidden units needed, as well as the optimal pooling strategy and feature type for the CNN and best results are achieved using large corpora (300-400 hours) and a two-convolutional-layer DNN with with 424 hidden units and four fully connected layers with 2,048 hidden units each, followed by a softmax layer with 512 output targets. This network yielded a 11.5% WER on that Switchboard corpus and 15% on the Broadcast news task with 50 hours of training data.

The most recent experiments involving CNN have included larger networks. The best CNN performance on TIMIT to date comes comparably close to Graves, Mohamed and Hinton (2013)at 18.2% PER which was obtained by Zhang *et al.* (2017) using a 10 layered CNN. Similar to the former study, the authors used the TIMIT training set and the development set to tune the parameters. Interestingly the authors comment that they believe that it is easy to overfit the TIMIT database due to its small size.

3.4 State of the Art using HTSVM for Speech Recognition

Hierarchical classification has not been often applied to the phoneme recognition task but some notable exceptions exist, like (DEKEL; KESHET; SINGER, 2004; KARPAGAVALLI; CHANDRA, 2015; DRIAUNYS; RUDŽIONIS; ŽVINYS, 2015) and (AMAMI; ELLOUZE, 2015). As a note, hierarchical classification does have multiple meanings, even within the field of machine learning. Here we assume this to be a hierarchical tree-like structure of classifiers which classify clusters of labels eventually filtering their results into the leaves which are represented as the individual and original class labels. A summary of the results presented in this section can be seen in Table 6

Table 6 – SotA Results for HTSVM Phoneme Recognition

Task	Best Study	Method	Size	WER	PER	FER	FA≤10ms
LD Phoneme Recognition	Driaunys, et. al.,2015	HTSVM	25k Phoneme Inst.	34.2	31.6	-	-
Tamil Phoneme Recognition	karpagavalli, et. al.,2015	HTSVM	2.4k Phoneme Inst.	-	33	-	-
TIMIT Phoneme Classification	Amami & Ellouze, 2015	HTSVM	2.4k	-	ca.40 -50%	-	-

While it does not implement SVM nor produce classification results based on any real corpus, Dekel, Keshet and Singer (2004) proposes an online algorithm based on techniques from large margin kernel methods and Bayesian analysis for phoneme classification and provide the theoretical background for their proposal, showing it to be better than a ”greedy” algorithm. In this case, the task of phoneme classification is treated as an optimization problem where a hierarchical tree structure divides groups of phonemes as nodes in the tree. The authors also found that the tree would tolerate small tree-induced errors while avoiding gross errors as a standard multi-class classifier would be prone to commit. More notable studies actually using this approach would follow in some years to come and the most notable studies within the last couple of years.

Recently, the HTSVM has been employed using data from speech corpora and applied this to a phoneme recognition task as presented in (DRIAUNYS; RUDŽIONIS; ŽVINYS, 2015), (KARPAGAVALLI; CHANDRA, 2015), and (AMAMI; ELLOUZE,

2015). Since this project also uses an HTSVM, I will sumarize each of these studies and their results.

In (DRIAUNYS; RUDŽIONIS; ŽVINYS, 2015), an experiment on stop and fricative consonants using the Lithuanian LTDIGITS corpus containing over 25,000 phoneme instances is presented. The most important findings were a 3% gain in the overall accuracy, a total of 68.4%, while reducing 52-55% of the computational time taken for classification with SVMs. The input features for the SVM were MFCC features. It is important to note that the corpus is made up only of spoken digits, making it a very restricted task.

In (KARPAGAVALLI; CHANDRA, 2015), a Tamil corpus of repeated words was developed and 2,400 phoneme instances were tested resulting in about 67% total accuracy on obstruent and sonorant sounds using MFCC features. Again one can notice a fairly high accuracy on a highly restricted task like identifying repeated words. It is difficult to know exactly how many unique words were used but the number of phoneme instances seems low for it to have been a very broad coverage. Also, once again MFCC features were used which are known for their practical use but also seem quite simple for a robust classifier like SVM.

In (AMAMI; ELLOUZE, 2015), a study on the TIMIT corpus presents MFCC classification results as well but more interestingly for each phoneme and major confusions classified by SVMs as well. Most of the phonemes fall between the accuracy range of 30% and 60%. The authors point out that due to the multiple dialects present in the TIMIT corpus, many phonemes are pronounced similarly to others depending on the speaker, increasing the confusion rate for similar phonemes where the SVM was not capable of efficiently classifying phonemes in the lower nodes of the tree. Unfortunately the overall classification is not given but it is likely to be somewhere below 50%. The level of classification is known to be a difficult task so it is not surprising that the results are lower than those in the repeated words and spoken digit experiments.

3.5 State of the Art in Acoustic Modeling

We will divide this topic in the three lines of research we have mentioned here in this project: (i) Forced Alignment (FA) and (ii) Phone Error Rate (PER), and related to PER (iii) Frame Error Rate (FER) will have their own sections. We want to draw special attention to the FA results since these experiments are not impacted by a language model, still they do use pronunciation models which are known to be essential for producing good results. For PER, the results should be useful; however, in most they are often skewed because they are calculated after a pronunciation and language model have been used. This makes PER and WER less reliable metrics for acoustic models. FER results tends to be the winner for transparency. While often some smoothing is used in post-processing,

as we use here, it is hard to mask the results of FER. On the other hand it is difficult to find top-tier studies which use this metric. Due to the lack of transparency of the relevant metrics, this section will perform a triage on them all and the careful reader can come to his own conclusions. The results will also be summarized is Table 7

Table 7 – SotA Results for AM using FA, PER, FER as metrics

Task	**Study**	**Method**	**Size**	**Rep?**	**PER**	**FER**	**FA≤ 10ms**
Forced Align. English Court Hearings	Yuan & Liberman, 2008	HVite	25.5 h	Y	-	-	**75.09**
Forced Align. for PRAAT	Goldman, 2011	HVite	30 min.	Y	-	-	50.5 English **51 French**
Forced Align. for BP	Souza & Neto, 2016	HVite	160 h	Y	-	-	**31.34**
TIMIT Benchmark Phoneme Recognition	Graves, et al., 2013	BLSTM RNN w/ PT	4 h	N	**17.7**	27.88	-
TIMIT Benchmark Frame Classification	Song & Cai, 2015	CNN+ CTC	4 h	N	29.4	**22.1**	-
TIMIT Benchmark Replication	van Niedek & et al., 2016	DLSTM-RNN Replication	4 h	Y	25.4	29.4	-

Table 7 shows the SotA results which were compiled, which use FER and PER metrics. In the table, each study is listed with the method used, whether the paper

provides sufficient information to reproduce it exactly (Rep? - short for reproducible) and the relevant metrics. It should be noted that the forced alignment studies will not have results for PER and FER and vice versa.

3.5.1 Studies Presenting Forced Alignment Results

Phonetic alignments can be generated in one of two ways: (i) manually, usually with the help of specialized software or (ii) automatically, as is done with FA. Manual alignment guarantees a high level of quality but it not free of subjectivity and is a high cost activity. Normally, at least two specialist transcribers are required to generate a kappa score or something similar and it is both time consuming and expensive as it requires at least two highly skilled transcribers and if one wants to annotate even a small database in a medium to long term time frame, it will be necessary to hire more. According to (SCHIEL *et al.*, 2012), manual transcription requires more than 13 hours/minute of audio. Often, researches opt for automatic solutions like FA. One the most used tools in the industry is HTK's HVite (YOUNG *et al.*, 2002). It is much more cost efficient to at least start with an automatic solution and then, if necessary, revise it by experts.

HVite is probably the most widely used tool and is the base software for the P2FA FA tool (YUAN; LIBERMAN, 2008), developed at the University of Pennsylvania. The authors show that their forced alignment can be as accurate as 75.09% for errors of less than or equal to 10ms and 93.92% for errors of less than or equal to 20ms for English. The corpus used was the 25 and a half hour long SCOTUS corpus with audio from 50 years of US supreme court hearings. The corpus features a good amount of reverb and multiple microphones but is fairly clean, making it a less than easy, but not overly difficult corpus to model. The authors use the CMU pronunciation dictionary(WEIDE, 2005) vowel stress to provide the training material (GMM-based monophones) for the forced aligner. The model is an HMM model and has 32 Gaussians Mixture components for each state with 39 PLP coefficients (12 Cepstral coefficients plus energy, Delta and Acceleration). They also correct a well known rounding issue in HTK which makes this tool very desirable to use. These results are interesting because they give some insight to the actual frame accuracy of these acoustic-pronunciation model systems and are both reliable and reproducible.

Another well known tool developed for this task is EasyAlign (GOLDMAN, 2011). This tool is also an HVite-based tool and has the advantage that it is built as a plugin for PRAAT (BOERSMA, 2006), which is by far the most widely used tool for manual revision of phonetic alignments. It was developed at the Université de Mons in France for French and English. French is well-known to be a rather difficult language for speech processing on the phonemic level. French includes many nasal sounds which make segmentations difficult and also many silent letters in its orthography making the decision to disambiguate the lexicon based on acoustics a far from trivial decision. The tool is able to make about 50.5%

of boundaries correct within 10ms for English and 51% for French. When the margin of error is widened to 20ms the accuracy rises to 76% for English and 80.5% for French. It is a bit less accurate than P2FA but was also trained on very little data being a 30min multi speaker corpus for each language and tested on 15 minute corpora for each language as well.

Recently, another tool has been developed for Brazilian Portuguese, which is interesting since this project focuses of non-native English speakers who are natives of Brazilian Portuguese. The tool, called Aligner (SOUZA; NETO, 2016), was developed by FalaBrasil, a group from UFPA (Universidade Federal do Pará). They used a rather large, unaligned audio corpus to train the triphone-state HMMs (trained with 12 MFCCs and C0 as the energy component). This is typical for HTK training. Like French, Portuguese is also considered a difficult language for phoneme segmentation and even experts often disagree on boundaries. The corpus used for training consisted of 160h of data and the test corpus was made up of hand-aligned test corpus made of 181 utterances recorded by a male speaker for a total duration of 7min. and 8s. The results were a bit lower than other studies at 31.34% for 10ms or less and 57.27% for 20ms or less. The authors did test the same acoustic models with EasyAligner and most results on any distance marker were within a percentage of point better or worse than EasyAligner. This gives the impression that the issue is probably not with the tool itself but rather the training material used. It may have been too poorly aligned or too noisy. Also it would probably have helped to use more test data.

3.5.2 Studies Presenting Phone Error Rate

PER is the industry standard for measuring the accuracy of acoustic models and is calculated by the the Levenshtein distance (LEVENSHTEIN, 1966) where the number of insertions, deletions and substitutions are added and divided by the total number of phonetic units in the string.

Often WER is calculated in the same way using the number of recognized words in a sentence; however, good WER results come from a good tandem of both acoustic and language models. This task is still far from resolved but has enjoyed great improvement in recent years. In 2009, Hifny and Renals (2009) was able to top the benchmark on the TIMIT corpus at the time with 23.0% PER using augmented context conditional random fields, besting the state-of-the-art HMM-based models at the time.

In 2013, Alan Graves benchmarked the TIMIT corpus at 17.7% PER, which seems to be the record that still stands according to (ZHANG *et al.*, 2017). This mark was achieved with Grave's Kaldi recipe, using a deep-bidirectional-LSTM RNN. This network included 3 Hidden layers with 250 units in each and pre-trained finite state transducers. It also used the 50 speaker development set for fine-tuning and early stopping. While

these results are impressive, they are not robust since they were highly fine-tuned for the TIMIT set. The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. This corpus will be described in further detail in the Methodology Chapter of this thesis.

Again, it is important to point out that the results presented here are reported for entire ASR systems (including pronunciation and language models) and not for the acoustic model in isolation. A Frame level analysis would make for an interesting compliment to these studies.

3.5.3 Studies Presenting Frame Error Rate

Some studies do report FER which is important for transparency in research when comparing acoustic models, because it represents the true accuracy of the acoustic model. FER is simply the accuracy of the phonemic labels attached to each frame used for training and testing. Typically, for speech processing, frames of 25ms are used with a stride of 10ms. This is logical because it is nearly impossible (with few exceptions, e.g., creaky voice in English or some taps in Portuguese) for phonemes to have a duration of less than 25ms and even if it is we are likely talking about missing just a few milliseconds which would still be unlikely to change the label and any segmentation error is unlikely to actually be perceived by the listener, even for TTS (especially in the case of parametric approaches). For speech recognition, in particular, this case is a non-issue. Beyond that fact, a good number to remember is 75ms which is the average duration of most of the smaller phonemes (mainly consonants). Divide this by three and you have three nice frames for your phoneme in general. In the case of longer durations, like vowels, the state simply will not change in the middle. So, in general 25ms is a good frame length for this task.

Of the few studies which present this metric a few benchmarks have been set. In 2013, Graves extended the work in Graves, Mohamed and Hinton (2013) with Graves, Jaitly and Mohamed (2013) where the same mark of 17.7% PER was achieved and additionally FER is included at 27.88% FER (both records at the time). Again the DBLSTM-RNN with 3 layers with 250 hidden units was pre-trained with CTC (Connectionist Temporal Classification). The 50 speaker development set for fine-tuning and early stopping as well as a biphone language model for predictions were used. The downside is that the study is not reproducible since it uses the pre-trained weights from the Google corpus.

On the note of reproducibility, (NIEDEK; HESKES; LEEUWEN, 2016) made an attempt to reproduce the network by (GRAVES; MOHAMED; HINTON, 2013). The main difference was that a DLSTM instead of a DBLSTM is used. The author explains that this was because of the lack of availability of a Bi-directional LSTM within the TensorFlow

library. Still, he does use a three layer LSTM network with 250 hidden units. The author used the openly available default initialization from TensorFlow based on (ZAREMBA; SUTSKEVER; VINYALS, 2014), since the data from (GRAVES; MOHAMED; HINTON, 2013) was not available. His network used a mini-batch size of 6 sub-sequences of 20 frames and applied dropout regularization. The final results of this reproducible network were 29.43% FER and 25.36% PER. In a second experiment he used Adam optimization with a learning rate of 0.001. The dropout regularization used a probability of 0.8. Here, he was able to better the results at 28.14% FER and 23.76% PER. The FER results are quite comparable to (GRAVES; JAITLY; MOHAMED, 2013) but the PER is much higher. This is probably due to the bi-phone language model. It also seems like the pretraining was rather similar using the TensorFlow default and whatever database was used in the previous studies. The author also provides his source code on his github repository. This kind of practice is crucial for the advancement of scientific research in this area. While the study does use default initializations from Tensorflow, at least all sources are available.

As far as FER is concerned, the best work is (SONG; CAI, 2015), beating the mark set by (GRAVES; JAITLY; MOHAMED, 2013). This network, to our delight, uses a CNN as well with CTC and achieves a FER of 22.1% but does not break the mark for PER. Again, this presents more evidence for the use of the bi-phone language model in (GRAVES; JAITLY; MOHAMED, 2013). Still, the authors do not offer any evidence for the generalization capacity but they do explain their training philosophy as stated in the article: "we train until the model begin[s] to overfit on the training set and the dev accuracy begins to fall. Much of the training is done on SAIL's Deep clusters, which uses nVidia GTX780 GPUs". The network with the best performance was the 25 frame windowed 128-256-384-384 CNN followed by 1024-512 dense layers. They do not; however, give the hyper-parameters used in their CNN and explain that they used their own pretrained language model for predictions and a pretrained RNN-CTC from some dataset (also not described). Since none of these resources are made available, this study is also not reproducible.

Apart from these benchmark studies, there have been a few other honorable mentions which also dive into other interesting topics, like classification windows. One such study is Lombart, Miguel and Lleida (2014). Here, a hybrid CNN and layer-fused MLP (Multi-layered Perceptron) with inputs of 11 frames of 25ms (step of 10ms) as context was used and trained with the TIMIT database. This study is interesting because it presents both PER and FER (Frame Error Rate) and because it also uses a CNN for feature extraction as we do but opts for and MLP-based classifier. In this paper, the best network had 512 neurons in three layers with the input layer consisting of 1024 neurons in the hidden layer and was able to obtain a FER of 43.04% and a 26.96% PER on the TIMIT test set. They also use articulatory features which is similar to the HTSVM strategy used in this thesis, at least in the spirit that articulatory features can be good indicators for

phoneme classification. They also use MFCCs and filter banks in the input for the CNN where this project uses the spectrogram image. Another detail is that they use 11 frames for classification. This is important as it allows the network to capture co-articulation. Little information is given about the CNN or how it was trained to extract the features.

Another study which uses larger frame windows for context is Lopes, Perdigão *et al.* (2009). In this study, a hierarchical broad-phoneme MLP/HMM hybrid classifier with MFCC features was used with window widening which achieved impressive results. For a 90ms window the FER was 61% using their past-future method that achieved 42% using 170ms and 17 frames for context training every other frame along the right and left side of the central frame for that period. It should be noted that these experiments were done with the full 61 phoneme set and results for smaller windows with 39 phonemes are not presented. It is interesting that the paper is from 2009 and few works seem to have presented better results since the boom of deep learning.

CHAPTER

4

METHODOLOGY

In this chapter, we will explain the approaches used in order to develop and evaluate the acoustic model proposed in this project. First, I will go back and review the research questions in Section 4.1, this time putting them into the context of the literature. Then we will start to get into the core resources of this project in Section 4.2, where I will go over the materials which were used: 1.) the speech corpus created during the project in Section 4.2.1; 2.) phonetic balancer script in Section 4.2.2; 3.) the interlingual pronunciation model, G2P converter and rule-based Brazilian pronunciation algorithm are presented in Section 4.2.3) as well as the datasets used for various experiments in 4.2.4, which will include the TIMIT dataset used for initial experiments in Section 4.2.4.1 and the processes for the dataset creation, especially forced alignment on TIMIT in Section 4.2.4.2 and the automatic segmentation and manual revision on the corpus produced for this thesis in Section 4.2.4.3.

After the data has been well described, we can begin going over the pipeline of the acoustic model in detail and step by step. This will begin in Section 4.3 with the features and ML algorithms used. Under that section several subsections will be included from the description of the spectrogram images in Section 4.3.1, the feature extraction process with the CNN in Section 4.3.2, the classification by HTSVM in Section 4.3.3 and finally the post-processing for PER used in Section 4.4.

To finish the chapter I will describe the process of data analysis and evaluation, both intrinsic and extrinsic which will be used at the end of this thesis in Section 4.4

4.1 Review of the Research Questions

Here we want to know how can the accuracy gap between native and non-native speech recognition be closed in an objective way. We will do this with two questions:

1. Does the model achieve results which are better than the SotA and bring us closer to the results produced by manual alignment?

   In the last chapter, the state-of-the-art research was established on the phoneme and frame level which are the metrics which we will use in this thesis. Speech recognition is known to produce poor results with non-native speakers (WANG; SCHULTZ; WAIBEL, 2003; VU *et al.*, 2014) and there is an added challenge when native speech recognition is not to be damaged as well. The closer the model can get to the numbers produced in studies like Graves, Jaitly and Mohamed (2013) and Abdel-Hamid *et al.* (2014) the better. For any supervised approach, a well annotated speech corpus is necessary to begin to close this gap. A speech corpus with both native and non-native speech is essential. While many strategies do exist, all of them search to optimize the resources they have in some way. When data and computational resources are available, many researches opt to capitalize on brute force strategies where domain and speech corpus adaptations could be considered a good cost/benefit. When resources are scarce, as in the scope of this thesis, the key is to optimize the quality of the data that is available. Here, this is treated as both a data quality issue as well as an architecture optimization issue.

   We assume that a good high end threshold for phoneme recognition is manual alignments, assuming that human segmentation is the gold standard used in most modern toolkits. This assumption is made because it eliminates any confusion generated by the language or pronunciation models. The closer the acoustic model comes to this mark, the closer it comes to perfection. The metrics used to evaluate this performance would be PER and FER. The PER shows us how well the model can recognize phonemes in a given utterance, which is the main task of the acoustic model. The FER is even more precise and shows how well the model can identify these classifications on the boundary level. This is important for research in automatic segmentation and could be considered an advancement in the state of the art.

2. Is the model robust?

   Robustness is treated here from a SLT perspective where convergence of the generalization on the validation set with a sufficient number of samples is shown on both the feature extraction level as well as the final classification. As explained in the theoretical foundations, the VC confidence and Chernoff bound will be used to approximate the empirical and expected risk of the CNN and SVM. While there is no perfect way to prove generalization without having all of the data in the universe, it does make an attempt to propose a model with much stronger confidence than those currently represented in the state-of-the-art literature.

Furthermore, metrics like FER and F1 are used on the validation sets for intrinsic evaluation of robustness and extrinsic evaluation, in this case a task-based evaluation, on a particular error will provide extrinsic evaluation. Also the method is validated on TIMIT, a native language benchmark, as well as the project corpus which is a interlingual corpus.

4.2 Resources and Datasets

The goal of this project is to build an acoustic model which accurately identifies native and non-native phonemes uttered by native English or native Brazilian speakers, recorded on home laptop computers. Since resources for tasks like this one are scarce, as evidenced in the table of existing speech corpora for non-native speech in the chapter on theoretical foundations, many secondary resources had to be developed in order to create the material needed for the acoustic model. These resources will also be made public and hopefully will make the life of researches who may follow this path a little bit easier.

This section will first go over the resources used to build the acoustic mode. The speech corpus created during the project will be detailed in Section 4.2.1, explaining the corpus design and annotation used as well as the phonetic balancer script developed for the project in Section 4.2.2. The interlingual pronunciation model will be covered in Section 4.2.3, explaining how it was created followed by an explanation of how it was augmented by the G2P converter and a script which creates an augmented dictionary from Brazilian accented English pronunciation rules.

The next parts of this section will detail data collection process in Section 4.2.1 and all of the specifications as well as the recording phase. Then, in Section 4.2.4, the datasets used will be described. First, I will start with TIMIT in Section 4.2.4.1, the standard native corpus used in the literature for this type of task and the forced alignment and all processing steps performed on TIMIT to get it to a point were it could be usable in Section 4.2.4.2. Last but not least I will explain in detail how the automatic segmentation and manual revision was carried out on the current corpus produced for this thesis in Section 4.2.4.3.

4.2.1 Ramble's Speech Corpus

The most essential item in this project is the speech corpus. In order to recognize non-native speech, a corpus including non-native speech must be utilized to train the acoustic model. In order to map the difference between non-native and native speech, in addition to non-native speech, native speech must also be trained in the acoustic model. Ideally, a speech corpus should be designed and recorded to reflect the environment and purpose for which it is to be used. In the current case, this specification would be a

typical study environment usual for students with common equipment readily available to them. Most students would use an online language tutor from their home laptops or computers, therefore, it would be best if the corpus was recorded on such equipment. A similar assertion goes for the environment where students could be expected to study in "semi-quiet" places like a dorm room, university house, coffee shop or family home. There could be some background noise but it should not overpower the voice of the person speaking. What is likely to generate more noise in this case is the actual recording itself, recorded by real users, i.e., people who have no background in audio recording. This means that it is expected to have some audio signals stronger and louder than others, clipping, equipment noise, etc.

The following instructions are given to any participant before recording:

"This corpus will be used to train an acoustic model for an English Pronunciation Tutor. Please record your sentences in a place where you would feel comfortable studying and with equipment that you would use to record your voice while using an online pronunciation tutor. Please note that no special equipment is required, only a laptop or computer and a standard built in, desktop or USB microphone."

The corpus is a phonetically balanced corpus and largely follows the algorithm presented in (MENDONÇA *et al.*, 2014) with some improvements to be described in Section 4.2.2. The phonetic transcriptions of the balanced sentences were then revised manually with a narrow phonetic transcription for both Brazilian and American speakers, all speaking English. The base corpus was created by rebalancing the corpus used in Almeida (2016), which was created by the other of that thesis together with the author of this dissertation. Then sentences were collected from the Wikipedia dumps in 2016 to create the final balanced corpus. Wikipedia was chosen because it is a readily available corpus which includes a wide variety of speech and an expanded lexicon. Wikipedia, as an encyclopedic corpus is a good cost/benefit and should be sufficient for the acoustic model. The alternative would be to collect a large amount (millions of sentences) from scientific conference presentations and thesis/dissertation defenses. This would simply be to costly and it would seem logical that a similarly well balanced corpus on scientific topics could be obtained from Wikipedia since many scientific topics from computer science, engineering, etc. can be found there. Also the language used in an encyclopedic corpus tends to have a wider range in coverage than scientific papers which was judged to be useful. Another benefit of this method is that large amounts of text could be easily extracted, facilitating the work for this thesis.

The data collection process was carried out online so that the users could record in their own environments. Overall, there is a great deal of noise in the corpus, including background noise like cars or heavy machinery, also at times other speakers are heard with various degrees of loudness, doors opening and closing, etc. The most damaging factors,

however, are usually due to microphone settings, too much or too little gain, muffled speech or just inferior microphone quality.

As the aim of the speech corpus is to be used in a pronunciation training system, it is transcribed with a narrow phonetic transcription (a transcription which includes allophones) explicitly designed to account for differences between the L1 and L2 used by the system, in this case English as L2 and Brazilian Portuguese as the L1. Currently, the corpus contains two parts. The first consists of 6h58m22s of speech being 1h23m20s from 13 (4 M/9 F) Americans and 5h35m02s from 49 (29 M/20 F) Brazilians. This part was rigorously annotated phonetically with manual revision of the phonetic alignment boundaries.

The second part contains a further 4h017m14s from 47 (23 M/24 F) Brazilians. This part is phonetically annotated and lightly revised speech with no manual segmentation revision; however it was forced aligned by an HTK-based acoustic model trained by the prior part and mainly features speech from from Brazilians. This brings the corpus to a total of 11h15m36s with 56 male and 55 female speakers.

The majority data was collected through the ICMC Institute. In some cases, my advisor and colleges from the NILC laboratory kindly allowed me to bower multiple laptops to set up a "recording stand" on campus. Most of the data was recorded by students from their homes and much of the "advertising" was done during English Academic Oral Presentation courses given by Dorly Piske on two occasions, in 2016 and 2017, where she not only allowed me to pass out information and explain my project to interested volunteers but also did some "pushing" to get more data for this project. This favorable atmosphere allowed for an efficient data collection process. In the case of the American data, the majority are students from Ohio, where I relied on my contacts from The Ohio State University and some are family member or contacts passed along from other people.

For the interested reader the final breakdown for education level (completed) in the Brazilian data was: 22% Post-Graduate Degree; 31% Bachelors Degree; and 47% Undergraduate students. For the American data, it was: 35% Post-Graduate Degree; 50% Bachelors Degree; and 14% Undergraduate students. In order to understand the native accents better, participants were asked were they lived for the better part of their life until adolescence. The data was organized by the Federal State in Brazil and the results for Brazilians were: 2% Bahia; 20% Minas Gerais; 6% Para; 14% Parana; 4 % Rio de Janeiro; 6% Rio Grande do Sul; 4% Santa Catarina; 2% Sergipe; and 41% Sao Paulo. The American data was collected in the same way and organized by US states: 14% Florida; 57% Ohio; 14% Pennsylvania; 7% Washington; and as an exception, 7% Ontario, Canada. The age breakdown for Brazilians was: 65% 18-25; 29% 26-35; and 6% over 35 with the oldest participant at 60 years of age. The age breakdown for the North Americans was: 14% 18-25; 64% 26-35; and 21% over 35 with the oldest participant at 63 years of age.

4.2.2 Phonetic Balancer Script

As described above the corpus was selected from the Wikipedia dumps. After cleaning (removing non-content material, sentencing, tokenization, normalization, removal of long or short sentences, etc.) the result was 17 million raw sentences. To facilitate the reader Table 8 provides some statistics from the corpus. It should be noted that the "*" denotes that the mode excludes outliers like single words since a part of the corpus includes only singe words.

Table 8 – Corpus statistics

Number of utterances	7,282
Number of unique sentences	6,000
Number of hours of audio	6.97
Number words	40,934
Average/mode* words/sentence	5.62/11
Max/min Number words/sentence	28/1
Number unique words	4,106
Number unique triphones	20,305

Then the sentences were transcribed using the G2P described in Section 4.2.3. Since the corpus needed to be phonetically balanced, a balancing script was developed during this project. The balanced corpus largely follows the algorithm presented in (MENDONÇA *et al.*, 2014) with some improvements. For example, weights were added to make fine-tuned adjustments for balancing and richness, i.e., a more zipf-like curve or an elevated tail. Also, log2 was applied to all of the scores to reduce the weight of triphones already accounted for in the balanced corpus. The current algorithm can be seen in Equation 4.1, where:

1. b = target corpus (big);
2. s = current corpus, which is being built (small);
3. x = candidate sentence;
4. f(x) = score of candidate sentence;
5. M = number of triphones in the sentence;
6. $N_{t,b}$ = number of triphone "t" in the corpus "b";
7. $N_{t,s}$ = number of triphone "t" in the corpus "s";
8. T = total occurrences of all triphones;
9. min = 0 if the triphone is not yet present in corpus "s" or 1 if the triphone is already present.

$$f(x) = \sum_{t=1}^{M} \left(\frac{log_2\left(\frac{N_{t,b}}{T_b}\right) - log_2\left(\frac{N_{t,s}}{T_s}\right)}{M} \right) + min \tag{4.1}$$

The final balanced corpus contains about 6,000 unique sentences developed in two separate phases of the project. The second was designed to compliment the former. Some optimizations were also made like the use of a trie for searching, data in the memory was minimized and an interactive shell was added for manual selection of sentences in addition to the automatic balancer[1]. This is particularly helpful for dirty or repetitive datasets. The final results can be seen in Figures 4.2.2 and 21. For the most part it is clear to see how the log2 function helped smooth the curve where the meat of the corpus is well represented, the most left end is not overly represented and the tail of the graph is significantly lifted. This produces a balanced graph which is similar to the original corpus in distribution and well-suited as a speech corpus. It should also be noted that the

[1] The tool can be downloaded from www.https://github.com/CShulby/Balancer-Scripts.

balanced corpus is presented in order of the phoneme frequency in the original corpus in order to evidence that the balancing was faithful even where some spikes and gaps can occur as opposed to perfect balancing.

Figure 20 – Original Histogram

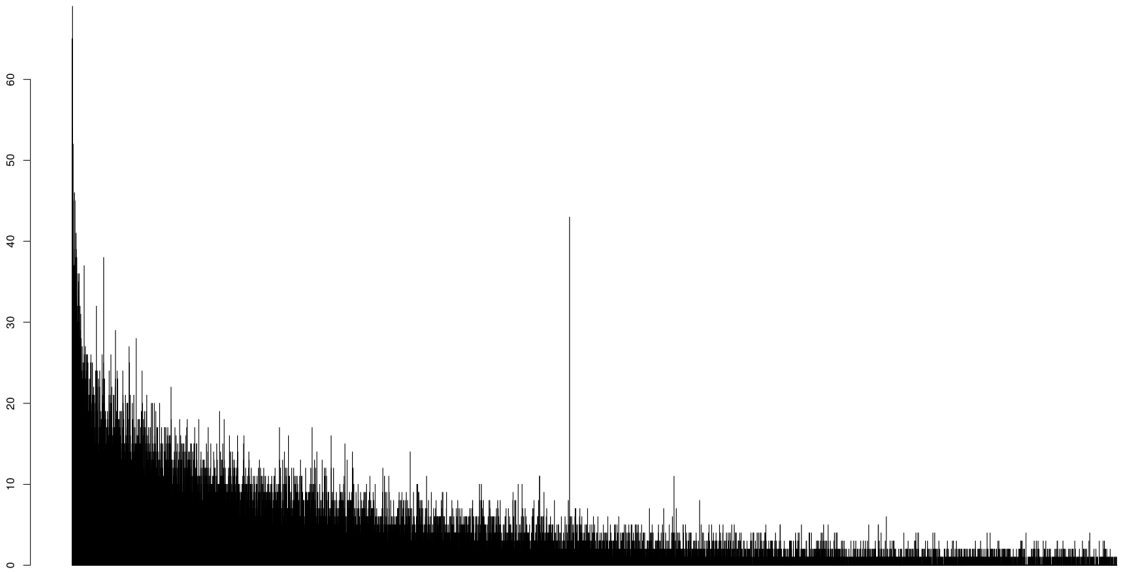

Figure 21 – Balanced Histogram

4.2.3 Pronunciation Model, G2P and Rule-based Pronunciation Algorithm

The project features an interlingual pronunciation model. All of the words in the corpus have been manually transcribed and included in the pronunciation model in order to train the acoustic model. This information is not only convenient, but also represents real utterances; it should be used first for recognition. Even though it is impossible to include every word that could be uttered in an academic conference, a good place to start would be by using a larger native pronunciation dictionary and via a rule based algorithm, expand that dictionary to account for the errors mentioned in Section 2.1. For the purpose of the project the seed dictionary is an adapted version of the CMU pronunciation dictionary (Weide, 2005). The adaptations are mainly in the phonetic notation to include BP-like phonemes.

Table 9 – Example excerpt from augmented PM

Word	Transcription
talked	t aa k t
talked	t ao k t
talked	t o k ih d
talked	t o k t
talking	t aa k ih ng
talking	t aa w k ihm
talking	t aa w k im
talking	t ao k ih ng
talking	t ao k iy ng
talking	t o k ihm
talking	t o k im
talking	t o k ih ng
talking	t o k im
talking	t o k iym
talks	t aa k s
talks	t aa w k s
talks	t ao k s
talks	t o k s
talk	t aa k
talk	t aa w k iy
talk	t ao k
talk	t o k
talk	t o k ih

Table 9 presents an example of the pronunciation model used. This pronunciation model differs from most ASR pronunciation models because it includes multiple variations of each word accounting for all predictable errors which are common for Brazilian learners of English as a foreign language. Combined with an acoustic model specifically trained with a narrowly transcribed speech corpus, it is easier to select the exact (at least as

close as possible within the predictable errors) transcription for each word; thus, allowing the system the resources to compare the actual input with the target input which could eventually be used in a tutoring system to present the correct feedback for the user.

The G2P converter uses a CART decision tree with pruning[2] to classify Out of Vocabulary (OOV) words. It was trained with an adapted version of the CMU pronunciation dictionary and forced aligned for unbalanced reference and hypothesis columns using the Many-to-Many aligner (JIAMPOJAMARN; KONDRAK; SHERIF, 2007) which is based on the stochastic transducer described in Ristad and Yianilos (1998). This aligner is meant to align letters to phonemes in a forced way. The output was then used to train the decision tree with heptagram padded windows. The G2P was used to convert new words to reasonable English transcriptions which were revised before being added in the dictionary. The G2P was implemented to facilitate the transcription of new words which were not present in the CMU dictionary. The G2P achieved a PER of 94.6% which was sufficient for this purpose. Still many pronunciations were edited manually for disfluencies. With a good pronunciation dictionary the initial HMM alignments for the training set required much less work when they were manually edited.

Once the pronunciations have been added into the lexicon, it is important to represent multiple possible variations of the word which a potential user could utter. A script[3] to augment the pronunciation dictionary was developed based on the errors mentioned in Section 2.1. For each of the nine errors, all phonological contexts prone to the error are duplicated, exhibiting the new pronunciation with the error.

4.2.4 Datasets

In this section, the datasets for different experiments will be covered. Initial experiments were carried out on the TIMIT corpus to establish an initial pipeline. The TIMIT corpus will be described in Section 4.2.4.1 and following in Section 4.2.4.2, the procedure for automatic and semi-automatic phoneme segmentation will be explained and will offer some anecdotes for future researchers. Then the automatic segmentation and manual revision procedures for the corpus produced for this thesis will be explained in Section 4.2.4.3, including the PRAAT plugin developed for this project (since integrated in the official CPrAN PRAAT repository) with the link to its source code.

4.2.4.1 TIMIT Corpus

The TIMIT Corpus, developed by Texas Instruments (TI) and the Massachusetts Institute of Technology (MIT) (GAROFOLO *et al.*, 1990), is a speech corpus meant to be

[2] from http://scikit-learn.org/stable/modules/tree.html
[3] The tool is available to download and use at: https://github.com/CShulby/augment_pron_-dictionary.

used for speech research and to serve as a standard for results comparison. TIMIT contains phonetically balanced prompted recordings of 2,342 unique sentences (2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1,890 phonetically-diverse sentences (SI)) from 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences to total 6,300 utterances (5.4 hours). The full training set contains 4,620 utterances. The test set contains 1,344 utterances from 168 speakers. With the exception of SA sentences which are usually excluded from tests, the training and test sets do not overlap and follow the suggested corpora splits outlined in (GAROFOLO *et al.*, 1990). TIMIT is considered a "balanced" corpus with respect to the distribution of phones and triphones found in the English language but it also follows a typical Zipf curve as one would expect to find in a speech corpus, where some phoneme sequences dominate a large portion of the samples and a large number of phonemes are less frequent. From a machine learning perspective, this is an unbalanced dataset. Normally, the phone set is collapsed into 39 monophones as suggested in (LEE; HON, 1989).

4.2.4.2 TIMIT Automatic Segmentation

In order to build a full pipeline, in this case with low-resourced languages in mind, the label generation was delegated to an automatic labeler, pre-trained with only one hour of data using a slightly modified version (small personal optimizations) of the P2FA Aligner (YUAN; LIBERMAN, 2008) which, as mentioned earlier, uses HTK's HVite tool.

To improve the performance of this tool, some additional adjustments were made, mainly a rule based script was created to generate multiple pronunciations for each word in the pronunciation dictionary as shown in Table 10. This was done to account for co-articulation and the eight dialects used in the TIMIT database and some light manual revision of difficult cases was done.

Table 10 – Example excerpt from the augmented pronunciation model

Word	**Transcription**
your	CH AO R
your	CH ER
your	CH OW R
your	CH UH R
your	JH AO R
your	JH ER
your	JH OW R
your	JH UH R
your	Y AO R
your	Y ER
your	Y OW R
your	Y UH R

Then the labeler was trained using the same tool on the rest of the TIMIT training

set. Then a final light revision was done for selected "difficult cases". As stated in the beginning of this section, this was done to simulate a situation where speech researches can not afford to do manual segmentation, as the TIMIT data is. Even if the researcher has a larger unannotated corpus (but maybe not hundreds of hours), this would also be a good strategy.

The main difference between this approach and the one used in the typical big data approach lies in the trade-off between knowledge and statistics. In a situation where we do not have enough data for a pure data-driven approach, it seems reasonable, from a cost/benefit perspective, to use a semi-automatic approach. Here this requires an investment of probably just a couple of weeks for a small team of annotators (2 or 3) to annotate an hour of speech from a balanced speech corpus. This should be enough data to generate decent forced alignment results (double what was used for easy align). Then, let's say the researcher has a remaining corpus of around 10-50 hours of speech. After the automatic alignment (which is likely to be somewhere around 60% for 10ms or less), the main work lies in refining only very difficult cases which is usually not a difficult task for an experienced annotator. In a few weeks and after retraining, the automatic aligner should be around the same accuracy as P2FA, if not better.

While it was not the focus of our work here, it suffices to say that I found the automatic aligner results to be exceptionally good with arguably little difference with the alignments provided in the TIMIT corpus.

4.2.4.3 Ramble's Automatic Segmentation and Manual Revision

In this subsection the automatic segmentation and manual revision procedures for the corpus produced for this thesis will be explained. It will be divided into two parts, first, the process for automatic segmentation will be described. Due to the noise in the corpus, this process needed to be modified from the original one set out earlier on the TIMIT corpus. Then the manual revision procedures will be described including the PRAAT plugin developed for this project (since integrated in the official CPrAN PRAAT repository) with the link to its source code.

Automatic Segmentation

A similar process for automatic segmentation as described in Section 4.2.4.2 was used. One hour of data was transcribed for initial alignments and the rest mas revised and retrained. The difference is that the revisions were made only in the orthographic and phonemic transcriptions. This is because even with one hour of data, unlike Timit, the corpus is simply too noisy to generate very good alignments. Instead, the model trained with one hour was used to automatically transcribe the rest of the corpus. OOV words were detected and included manually in the pronunciation dictionary and the sentences

were revised orthographically so that all prompts matched the words uttered in the audio. Of course it is impossible to account for all of the accented pronunciations but the idea was that the majority should be ok and errors would be punctual. This indeed was true and the corpus was revised by listening to the prompts and correcting the automatic transcriptions which were used in turn to expand the pronunciation dictionary. This pronunciation dictionary was then used to once again generate better automatic transcriptions which were used to train a new acoustic model with all of the training data. This generated much better alignments as well but still, due to noise, many issues were present and not all transcriptions were perfect. The next step is detailed in the following part, describing the process of manual revision carried out on the corpus with the HTKlabel Plugin.

HTKlabel Plugin for PRAAT

When annotating a speech corpus, one generally encounters two repetitive issues. The first is that speakers do not always say what is in the prompt. This happens for a variety of reasons, for example: the speaker misread the prompt; the speaker did not understand the prompt; slips of the tongue; external factors (the speaker coughs in the middle, is interrupted, etc.). The other issue is even more prevalent for non-native speech. Having the correct word in the dictionary does not mean that one also has the correct transcription for every speaker. In the case of accented speech this task can easily turn into more trouble than it is worth.

Here is an example of a prompt from the corpus:

It really gets on nerves|## ih^t. r iy^. l iy # g eh^t s # aa^n # m ay^# n er^v z ##

In a perfect world for a speech engineer everyone would say exactly what is written on the prompt with the same pronunciation. Unfortunately, it isn't a perfect world.

Here is an example of an actual utterance of the prompt above, transcribed:

It really gets on my my nerves|## ih^ch # ih. r iy^. l ih # g eh^s # aam^# m ah^# m ay^# n er^v s ##

In order to speed up the editing process a plugin for PRAAT was developed[4]. A plugin for Praat to read information from HTK and HTS label files (.lab) and Master Label Files (.mlf) and transforms them into their TextGrid equivalents for manual editing in PRAAT and then allows the user to save them again, back in their original formats. It is also part of the htklabel CPrAN (an official PRAAT repository) plugin for PRAAT. The htklabel plugin is free software, so one can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation,

[4] The latest stable version is available through CPrAN or at http://cpran.net/plugins/htklabel or the bleeding edge version can be found at https://github.com/CShulby/plugin_htklabel

version 3.

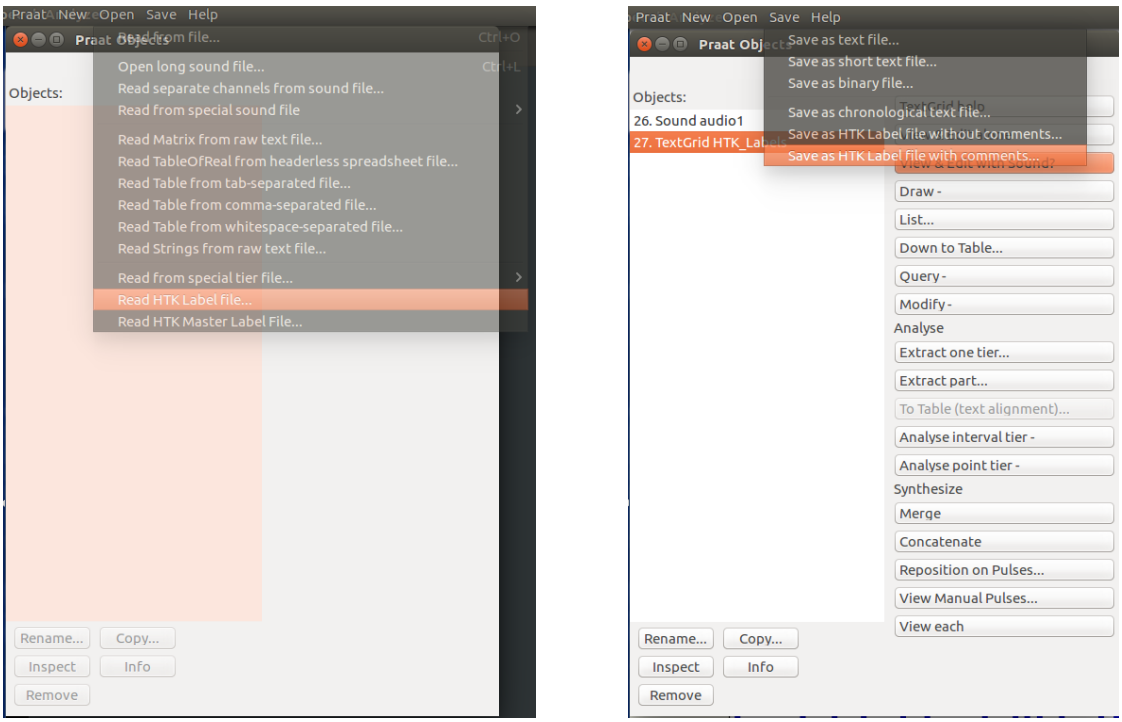

Figure 22 – The user screens in htklabel

In Figure 22, one can see the two step processes for opening and saving .lab files withing PRAAT via htklabel. In the first step one can open with: Open -> Read HTK Label file... and to save one can follow: Save -> Save as HTK Label file with comments...

4.3 Features and ML Algorithms

In the following sections the HTSVM architecture used in this work will be explained step by step from the feature extraction and classification to final post-processing procedures. One of the goals of this project is to be as transparent as possible in order to better understand what is required to build robust acoustic models. The full pipeline, which will be explained in detail throughout this section, can be seen in Figure 23.

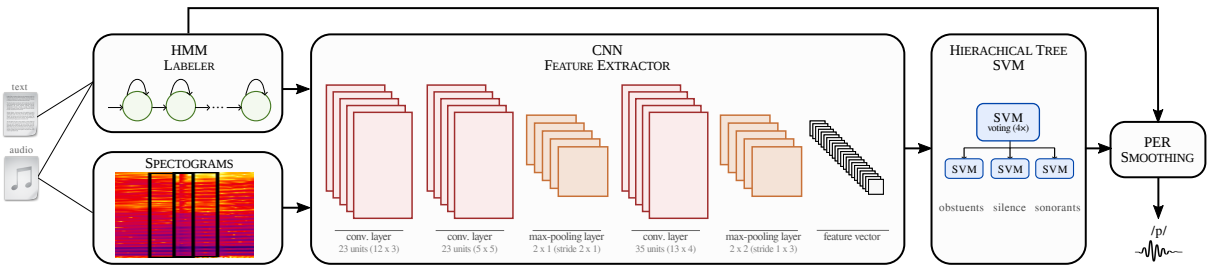

Figure 23 – The final pipeline used for Acoustic Modeling

4.3.1 Spectrogram Images

For each audio file, spectrograms of 25ms Hann windows with a stride of 10ms were extracted. This is the industry standard (MACLEAN, 2018; VERTANEN, 2018; POVEY *et al.*, 2011; GRAVES; JAITLY, 2014) and since some taps in this corpus do reach 20ms, it did not make any sense to use larger windows. This means that each window consists of 400 samples (25ms $\times$16kHz) and has a DFT of 254. The images were resized to 5×128 pixel images. This was done in order to reduce the number of features extracted to a manageable dimensionality while prioritizing the frequency information (which is simply reduced in the number of pixels equally over the entire image). These measurements were found after empirical tests. First, it was established that the images could be very narrow, then the height was tested empirically until the final size was found. When a sentence ended in a number of milliseconds which is not divisible by 25, the last frame was squeezed into the penultimate frame. Since there is always a little bit of buffer before and after the speaker speaks, in all cases, this last frame was silence as expected. In the case of TIMIT, this process yielded $1,447,869$ images for training and $482,623$ for testing. For the target corpus of this thesis, it was $1,447,869$ images for training and $482,623$ for testing

4.3.2 Feature Extraction

CNN is a deep learning technique, which presents good results in several domains (LECUN; BENGIO, 1995), including speech processing (ABDEL-HAMID *et al.*, 2012; SAINATH *et al.*, 2013; ABDEL-HAMID *et al.*, 2014).

As mentioned earlier, the biological motivations prompted the use of a CNN to extract relevant features from the spectrogram images, conserving only the most important information. The advantage of the CNN in this application is the identification of local recurrence information and its invariant translation in data (LECUN *et al.*, 1998; LECUN; BENGIO; HINTON, 2015; GOODFELLOW YOSHUA BENGIO, 2016). CNN are tolerant to distortion as they combine local receptive fields, shared weights and spatial sub-sampling. All three of which are useful in phoneme recognition (ABDEL-HAMID *et al.*, 2012). The trick comes in balancing the spatial resolution reduction with the representational richness of the images in order to generate the most useful feature maps at a low classification cost with high accuracy (LECUN *et al.*, 1998). We perform this in two ways: first, by rescaling the image to a smaller size while still preserving the most important information for phoneme classification and, secondly, by searching for the best sized masks. It should be noted that we use only single frames for classification.

The feature maps were optimized by searching for the best sized masks. The size of the masks and number of neurons were estimated according to the adapted FNN algorithm proposed in (FERREIRA *et al.*, 2018). The best five configurations found by the algorithm were generated and then selected the one with the largest mask size and fewest

neurons. This selection was mostly systematic although the decision between the final configurations was slightly subjective as we preferred larger mask sizes to fewer neurons where the difference was small. This "feeling" was guided by the logic that the largest avatar which best represents the data is the best choice for strong generalizations. It is interesting to admit to the reader that while empirical tests lead to the final selection, these tests were guided by intuition and confirmed the same beliefs used in feature engineering approaches. If one analyzes the filter sizes closely, one can note that with 12 pixels of height in the first filter and a 128 pixel tall images, just over 10 filter bands are created as input to the network. This is interesting since often nine to thirteen MFCCs are used in traditional approaches for ASR. The width of 3 pixels in the filter can be seen as more symbolic than anything since we do not analyze the time window in great detail but since we have a 5 pixel width, it makes sense to split this into overlapping halves to guarantee good feature extraction.

The final network is composed of 3 layers:

1. 23 convolutional units with a mask size of 12×3; followed by
2. 23 convolutional units with a mask size of 5×5 and a max-pooling of 2×1 with a stride of 2×1; and finally
3. 35 convolutional units with a mask size of 14×4 and a max-pooling of 2×2 with a stride of 1×3.

ReLU (Rectified Linear Unit) was applied as the activation function due to its widespread adoption in the literature. This function avoids negative values and maintains the scale of output values. The CNN was trained using the Keras (CHOLLET *et al.*, 2015) package developed with the TensorFlow library. The 988 input dimensions were generated using half or "same" padding as $(ceiling\frac{128\times5}{5\times5}) \times 38 = 988$. Where 128×5 is the size of the input images, 5×5 is the pooling size and 38 is the number of neurons used. The network was initialized randomly with no pre-training. Since SLT gives us sufficient confidence that we have enough training samples, no optimizations should be necessary. Observe that the half padding always rounds up using the ceiling function. It took about a day to execute the network using a single Titan-X GPU and 32GB of RAM.

4.3.3 Classification

The proposed method takes advantage of the machine learning techniques used for CNN and SVM and attempts to improve accuracy and minimize their disadvantages with a knowledge-based hierarchical tree structure.

The features produced by the CNN were classified using a SVM since it has been used in literature combined with CNN that has provided good results in many

domains (LECUN; BENGIO, 1995), including speech processing (ABDEL-HAMID *et al.*, 2012; SAINATH *et al.*, 2013; ABDEL-HAMID *et al.*, 2014). In addition, SVM provides a strong learning guarantee according to SLT and large-margin bounds (VAPNIK, 2013; LUXBURG; SCHÖLKOPF, 2008). The SVM parameters were found empirically after several experiments.

When selecting the kernel, multiple kernel types were tested with a default value of $coef0 = 1$ (as a non-homogeneous kernel) and a cost $C = 10$, was used on a set of 10,000 randomly selected images. Tests were carried out using cross-fold validation with 20% used for validation. The average results for the ten folds are presented in Table 11. Finally a grid search was performed to find the best value of C. The final configuration achieved an accuracy of 77.52% using the same set of 10,000 images.

Table 11 – Cross-fold validation results for SVM kernel selection simulations

Kernel type	**Accuracy**	**Standard deviation**
Linear	62.8%	1.32
Radial basis function	69.32%	0.85
Polynomial-2nd°	72.38%	1.1
Polynomial-3rd°	74.99%	1.86
Polynomial-4th°	76.07%	1.34
Polynomial-5th°	74.5%	1.8

The selected kernel for final experiments was a 4^{th} order polynomial kernel with $coef0 = 1$ (as a non-homogeneous kernel) and a cost $C = 10,000$. As in other studies on natural language processing problems, like Chang *et al.* (2010) and Goldberg and Elhadad (2008), we found a polynomial kernel to be more useful for this task than a RBF kernel. This is not surprising since the Zipf-like class distributions for these tasks are similar. For this task, namely speech recognition, we have two main problems to solve. First, the number of CNN extracted features combined with the number sample frames, since the SVM training time increases quadratically as the number of examples increases. The second problem is the unbalanced nature of the dataset, a common characteristic for most speech corpora.

For the first issue, the hierarchical structure is what makes the SVM a viable option. The cost of training a sequential SVM on this type of dataset would be prohibitive and even with a great deal of work in data reduction techniques, it would still likely take several months to train the model. By dividing the task into several hierarchical levels based on the knowledge of articulatory phonetic classifications in English as described in (LADEFOGED; DISNER, 2012), we are able to turn the problem into a binary, ternary or quaternary classification instead of the original 57 classes. The table with all of the classes used for this project can be seen in Table 12

Table 12 – Phoneme Classes

Phoneme	**Example**	**Phoneme**	**Example**
aa	father	iy	feet
aam	andar*	iym	team@
aar	farther	jh	jack
ae	cat	k	karen
aem	camera@	l	lip
ah	putt	m	mitten
ahm	pun@	n	not
ao	hawk	ng	morning
aw	cow	o	hawk@
awm	frown@	om	home@
ay	eye	or	fort@
aym	I'm@	ow	no
b	better	owm	nominal@
ch	charlie	oy	boy
d	dog	p	pet
dh	there	r	rex
eh	red	s	send
ehm	emily@	sh	shut
er	fur	t	tex
ey	pay	th	think
eym	pain@	uh	put
f	fit	uw	food
g	git	uwm	tune@
hh	happy	v	vera
i	/i/street@	w	wine
ih	bit	y	you
ihm	tim@	z	zoo
im	in@	zh	beige

Ladefoged suggests a hierarchical structure necessary for English features in the last chapter of (LADEFOGED; DISNER, 2012). This served as the primary source for our tree which was derived as possible questions to classify each phoneme so that they contain the features necessary for classification. This makes our classification space much more simple when creating the support vectors. It should be noted that the first layer classifying obstruents, silence and sonorants is built using 4 individually trained SVMs on equal chunks of data where the prediction for this layer is made by a simple voting system where the mode is taken as the final prediction. This was done to further reduce the training time. The hierarchical structure is presented in Figure 24.

In the second issue, we deal with an unbalanced dataset where even minimal pairs can have a large difference in frequency as in the example of /k/ and /g/ which are both velar stops. In the training set, the phoneme /k/ appears in $60,433$ frames, whereas /g/ is found in only $17,727$. In order to build a robust system, it is important to learn this

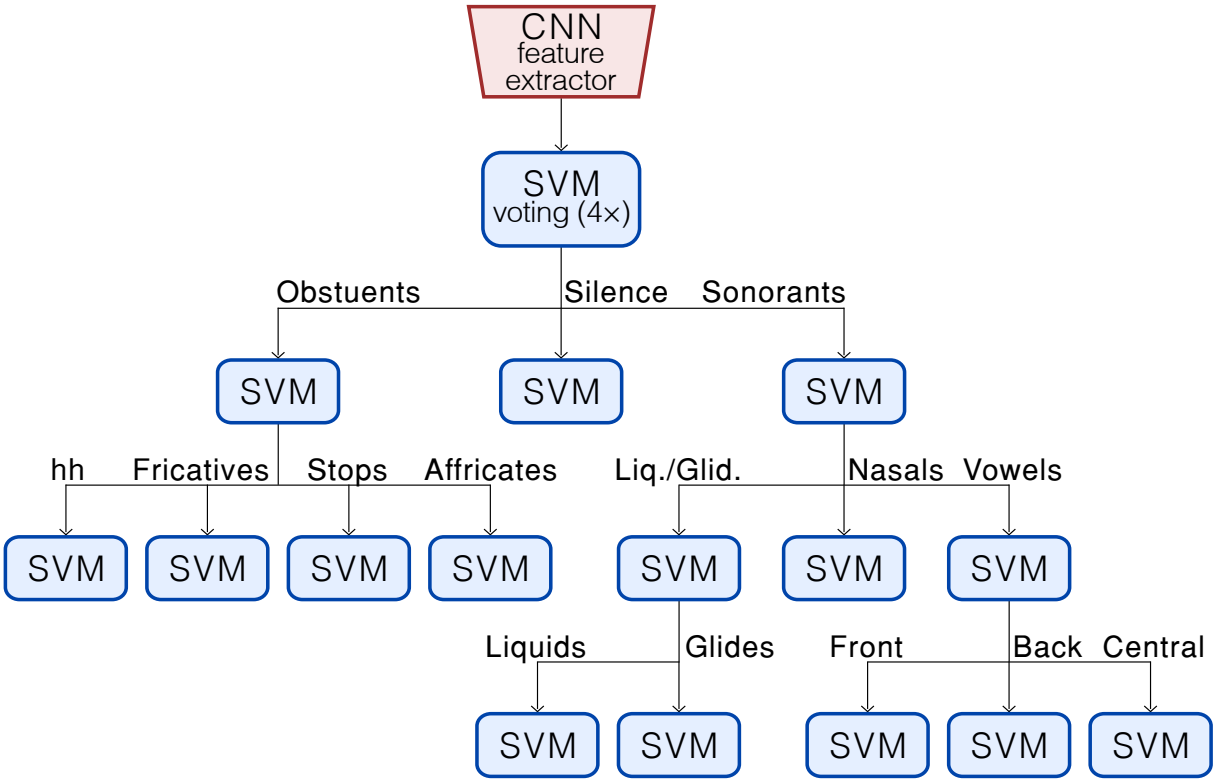

Figure 24 – HTSVM architecture defined for the experiments.

phonemic distinction and minimize the influence of probability in the training set.

The technique most prevalent in the literature for accented speech is the use of adaptive algorithms like MAP (WANG; SCHULTZ; WAIBEL, 2003; VU *et al.*, 2014). While this technique has proven effective, it doesn't solve the issue of generalization capacity, but rather simply masks the problem and for a specific dataset or task will generalize well, but not for others. For this reason, I have chosen to avoid such aggressive statistical adaptation techniques. More recently for multiple domains and noise variation Multi-conditional Training (MCT) has also become popular (KIM *et al.*, 2016). This is actually a bit more interesting from the generalization standpoint but it seems to fall into the "use a bazooka to kill a fly", brute force type method. Since the current proposal is more focused on low resource situations and more simple, thoughtful techniques, this type of method is out of the scope of this project. Therefore, since the idea is to minimize any large-scale statistical adaptations, the proposed method seeks to use a simple over-sampling technique.

While there are a number of methods available for data over-sampling unbalanced datasets, one of the most common, most simple and most useful technique is the SMOTE (CHAWLA *et al.*, 2002) data augmentation technique as seen in Figure 25 in comparison with other common over-sampling methods. In Figure 25, one can see that SMOTE is a robust over-sampling technique and well suited for a decision algorithm like an SVM, while other methods may be more useful for gradient-based methods. The

technique works as follows:

1. starting with training set which has s samples, and f features in the feature space;
2. to oversample, we take a sample from the dataset and consider its k nearest neighbors in the feature space;
3. to create a synthetic data point, take the vector between one of those k neighbors and the current data point;
4. multiply this vector by a random number x which lies between 0, and 1;
5. add this to the current data point to create the new, synthetic data point.

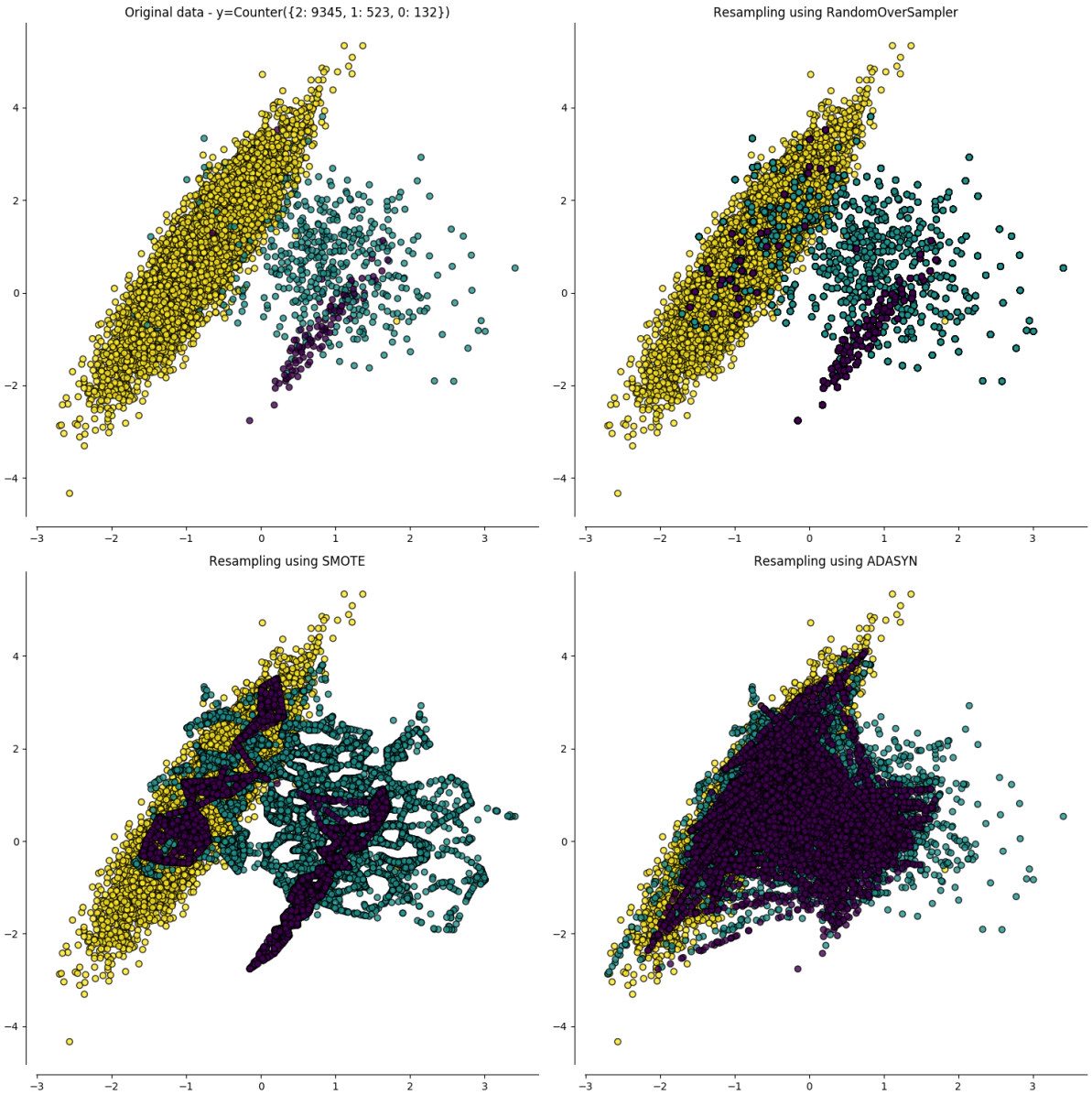

Figure 25 – Comparison of over-sampling techniques, imbalanced-learn 0.3.0

Synthetic creation of minority samples allows us to treat the classification task with more confidence. This was also made possible by the hierarchical approach because each node has a much smaller number of samples than the original dataset.

4.3.4 PER smoothing

In order to obtain PER, the classified frames must be converted to phonemes. This was done by taking the mode of all of the SVM classifications for each GMM-HMM generated boundary, thus collapsing repeated frame classifications into single phonemes. In the case where classifications across boundaries produced the same results, that boundary ceased to exist. Smoothing is an important step because it avoids likely misclassification from the frame level, for example: a single frame classified as a phoneme A between two frames classified as a phoneme B seems extremely unlikely.

4.4 Data Analysis and Evaluation

Intrinsic evaluation is designed to test the acoustic model itself, isolated from its practical purpose. In other words, in an intrinsic evaluation, the main focus is to evaluate the efficiency recognizer is able to obtain. Given an acoustic signal, its textual counterpart is compared for its reference and hypothesized output. The metrics used are FER, PER and F1 to measure the accuracy of the model.

The extrinsic evaluation is that which evaluates the purpose for which the acoustic model was built in the first place, in this case a possible application would be pronunciation tutor, the ultimate goal is the acquisition of properly pronounced words by the users of the system. This is why the feedback accuracy will be evaluated for a specific error, in this case, the most common and most communicatively impeding error, word final epenthesis. This task was chosen because it is a particularly difficult construction for Brazilians (JOHN; CARDOSO, 2017) and an error which greatly impedes communication.

CHAPTER

5

EXPERIMENTS, RESULTS AND DISCUSSION

In this chapter we will go through the several experimental steps throughout this project. Some important considerations first: the CNN-HTSVM approach is same for every experiment - best cost/benefit, careful attention to learning guarantees; the architecture changes based on the dataset but the method for building the architecture does not. For example: the mask, max-pooling, and convolutional unit sizes may change from TIMIT to Ramble corpus but they were both obtained by taking the three best FNN results where the smallest network, with the biggest mask was chosen every time. Some refinements are made along the way like CNN layers and classification context size to show how they impact results. This is done for experimental purposes but the final architecture is always the same.

The experiments covered here are 1.) Shallow CNN-HTSVM with convergence analysis in Section 5.1; 2.) Deep CNN-HTSVM with window-widening in Section 5.2; and 3.) the final CNN-HTSVM for non-native speech in Section 5.3.

5.1 Shallow CNN-HTSVM with convergence analysis

Three experiments were carried out in this section: 1.) GMM-HMM HTK model with 31 Gaussians; 2.) MLP with 1 Hidden Layer of 100 units and a ReLU activation function and 3.) a Shallow-CNN-HTSVM to show what the most simple architecture in this scope in capable of. These classifiers will be detailed a little further now. All models trained on the same TIMIT training set with a zero-gram Language Model (LM) to obtain only posterior values.

As a baseline comparison to the proposed method, one of the most popular ASR toolkits for a database the size of TIMIT was used, the HTK toolkit (YOUNG *et al.*,

2002). A triphone HTK model was trained on the same TIMIT training set used in the current method and recognition was performed on the test set with a zero-gram language model, with only the individual monophones as the pronunciation model. This was done in order to obtain only the posterior values from the acoustic model predictions without the influence of a language or pronunciation model for fair comparison, since only the accuracy of the acoustic model is being evaluated. 31 Gaussians were used because this number have been found useful in the past. It is higher than what is recommended by the voxforge tutorial at (MACLEAN, 2018), which uses 15 and is similar to the models used by Keith Vertanen at (VERTANEN, 2018), where he uses a maximum of 32 for the Wall Street Journal and TIMIT datasets together. The model was trained using MFCC 0DANZ acoustic features, where 0 uses the zeroth cepstral coefficient, D is for the delta coefficients, A for acceleration coefficients, computed as delta of delta coefficients, N is for absolute energy suppression and Z is zero mean normalization. The predictions were then segmented in the same fashion as the proposed method with 25ms sliding windows and a step of 10ms, in order to make a frame by frame comparison.

For feature extraction to be used by the MLP and the HTSVM, the shallow CNN architecture used in the experiments was estimated according to the adapted FNN proposed in (FERREIRA *et al.*, 2018) and is composed of 38 convolutional units with a 29×1 mask size. In the case of the sub-sampling layer, max-pooling was used with 38 units and sized at 5×5, without overlapping. ReLU was applied as the activation function due to its widespread adoption in literature. This function avoids negative values and maintains the scale of output values.

The second experiment was a simple Multi-layer Perceptron with one hidden layer and 100 neurons, a ReLu activation function and an Adam solver. The choice for 100 neurons was made because it does not saturate the network. Some experiments were done to confirm this logic with larger configurations like 1000 neurons but no more than an overall accuracy gain of 0.6% was obtained and the and loss did not improve at 1.992. After running multiple experiments, it seems that this was due to chance since some experiments were slightly lower and some higher. Deeper architectures with 2 or 3 layers presented worse results of about a 5% to 8% drop in accuracy over several experiments. This seems to be due to the trade off between complexity and generalization of the network and is an expected result due to the small dataset. The learning rate and all other hyperparameters used the default values from the MLPClassifier from (PEDREGOSA *et al.*, 2011) which was heavily based on the work in (HINTON, 1989). The network is also very similar to the one used in (LOPES; PERDIGÃO *et al.*, 2009). Again, for consistency, we chose to use a strong baseline network without hyper-parameter tuning. This was done to show the gain provided by the HTSVM structure over another widely used classifier for the classification of CNN features.

Finally, the The HTSVM described in the methodology chapter was used being: a 4^{th} order polynomial kernel with $coef0 = 1$ (as a non-homogeneous kernel) with a cost $C = 10,000$. As mentioned earlier in this thesis the SVM provides a strong learning guarantee according to SLT and large-margin bounds (VAPNIK, 2013; LUXBURG; SCHÖLKOPF, 2008). Also, the HTSVM tree structure was able to deal with the parallelization and unbalanced dataset issues.

5.1.1 Results

The idea behind these experiments was to compare a typical GMM-HMM model which is the state of the art for this type of task and an MLP classifier which is often used in the place of SVM in the literature with the current HTSVM approach. This is why the CNN was kept to just one layer and one frame to reduce influence by other factors and to streamline the analysis. Table 13 shows the F1 scores and FER of the GMM-HMM, CNN-MLP and CNN-HTSVM models for frame classifications as well as the PER in the case of phoneme classification. The FER was calculated as an accuracy score, PER was calculated by the conventional Levenshtein edit distance (LEVENSHTEIN, 1966) and the F1 score was calculated as $2 \times \frac{precision \times recall}{precision + recall}$.

Table 13 – F1 Scores in Frames, Frame Error Rates and Phone Error Rates for each model.

Classifier	F1 Score	FER%	PER%
GMM-HMM	0.166	76.36	75.17
CNN-MLP	0.225	56.97	52.90
CNN-HTSVM	**0.491**	**37.04**	**35.41**

The HTSVM presents strong results, even using only a single frame for classification, when compared with other studies like Karpagavalli and Chandra (2015), who use 9 frames for classification on a much easier task (spoken words). In that study FER is not given but a 33% PER is cited, less than 2.5 points better than this dissertations results on TIMIT. Lombart, Miguel and Lleida (2014) used 11 frames as context for classification and achieved a FER of 43.04%, 5 points worse than the results here, but a PER of 26.96%, which is a substantial 8 point difference. This study is interesting but it is also tested on the TIMIT dataset.

Independent of the models accuracy, it is also important to understand what sort of errors the models are actually committing. Table 14 lists the 15 most frequent errors committed by each system, including the true values, predicted values and the confusion percentage. The GMM-HMM systems are known to produce rather jumbled posterior values where they typically rely on providing a ranking to the pronunciation model where

these issues are normally solved. Still, this is not an adequate approach for phoneme recognition.

Table 14 – Most Frequent FER Confusion percentages

GMM-HMM			CNN-MLP			CNN-HTSVM		
True	Pred	Conf (%)	True	Pred	Conf (%)	True	Pred	Conf (%)
s	z	33.14	s	f	21.48	s	z	15.16
ih	uw	16.00	ih	ae	16.23	ay	ae	39.64
t	ch	17.58	iy	ae	14.99	ao	aa	26.58
er	r	32.46	z	f	28.01	r	er	18.84
ao	l	28.00	ay	ae	24.59	sh	s	26.01
iy	y	14.23	eh	ae	25.46	aa	ae	16.07
s	sh	10.09	aa	ae	20.96	ah	ae	14.79
ae	t	14.32	er	ae	14.32	t	s	7.61
ih	z	10.07	k	t	13.22	iy	ih	6.48
w	ao	45.52	ey	ae	25.29	er	r	7.64

Here one can see that the findings of Amami and Ellouze (2015) can be confirmed that the SVM does get confused when phonemes are very similar. Still, this may be caused by the transcription of the 8 dialects where some sounds, especially vowel sounds can have some overlap. One observation between the SVM and MLP classifiers is that the MLP makes more repetitive errors. Since the SVM is governed by a decision boundary function and the MLP is a probability of an argmax function, this makes sense. Even though the same CNN features were extracted, it seems that the MLP was more likely to develop a "trash" category where when in doubt it goes with a "more probable" class. In the example of /ae/, it was correctly classified 60% of the time so it became a go-to class when in doubt. In the case of /f/, the phoneme was classified correctly in 85% of the instances, so other fricatives were more likely to be classified as /f/ as well. The GMM-HMM is notoriously unsuccessful on the phoneme level which is why engines used tied states to calculate the cost of substitutions to find the correct transcription in the pronunciation model. What is most evident is that often when it errs, it fails badly. The MLP is also not the most graceful failure, where the SVM is more robust in that its failures are more reasonable as pointed out in Amami and Ellouze (2015).

5.1.2 Convergence Analysis on TIMIT

In this section the following items will be discussed: 1.) the motivation for performing a convergence analysis; risk assessment; 2.) generalization bounds and the VC theory applied; 3.) the convergence curves for the CNN and SVM will be shown and discussed.

The motivation for convergence analysis comes from a need for providing evidence for more robust acoustic models and machine learning models in general. Since the deep learning boom in the last half decade, machine learners have focused more on the application of these algorithms and their intrinsic results, rather than studying how well their models actually work in real situations. This is in part due to the difficulty of finding usage cases, but we can still provide statistical guarantees with a little bit of care and attention. These days, the SVM is well known for its robustness in SLT literature (VAPNIK, 1998; LUXBURG; SCHÖLKOPF, 2008). Therefore, having an SVM in the pipeline seems like a correct choice a priori.

SLT provides theoretical support for such convergence proofs in terms of how supervised learning algorithms generalize examples. Equation 5.1 defines the main principle of SLT which is the empirical risk minimization (LUXBURG; SCHÖLKOPF, 2008) to bound the divergence ε between the empirical risk R_{emp}, i.e., the error measured in a sample, and the expected risk $R(f)$, i.e., the expected error while assessing the joint probability distribution of examples and their respective classes, as the sample size n tends towards infinity. Still describing the equation, the right-most term is known as the Chernoff bound, f is a given classifier, and $\mathscr{F}$ is the space of admissible functions provided by some supervised algorithm, a.k.a. the algorithm bias (VAPNIK, 2013; LUXBURG; SCHÖLKOPF, 2008; Fernandes de Mello; Dais Ferreira; Antonelli Ponti, 2017).

$$P\left(\sup_{f\in\mathscr{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq 2e^{-n\varepsilon^2/4} \tag{5.1}$$

(VAPNIK, 2013) proved a bound for supervised learning algorithms considering the shattering coefficient $\mathscr{N}(\mathscr{F},2n)$, as defined in Equation 5.2. Such a coefficient is a measure function to compute the complexity of the algorithm bias, i.e., the cardinality of functions contained in the space $\mathscr{F}$ that produce different classification outputs, provided a sample size n. Throughout the formulation, the generalization bound defined in Equation 5.4 is employed, a further result obtained from Equation 5.2, to ensure that the expected risk is bounded by the empirical risk plus an additional term associated with the shattering coefficient and some probability δ (Equation 5.3).

$$P\left(\sup_{f\in\mathscr{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq 2\mathscr{N}(\mathscr{F},2n)e^{-n\varepsilon^2/4} \tag{5.2}$$

$$\begin{gathered}P\left(\sup_{f\in\mathscr{F}}\left|R(f)-R_{emp}(f)\right|\geq\varepsilon\right)\leq\delta \\ \delta=2\mathscr{N}(\mathscr{F},2n)e^{-n\varepsilon^2/4}\end{gathered} \tag{5.3}$$

$$R(f) \leq R_{emp}(f) + \sqrt{4/n(\log(2\mathcal{N}(\mathcal{F},n)) - \log(\delta))} \tag{5.4}$$

In the case of the SVM, the same bound is formulated as shown in Equation 5.5, in which c is some constant, R corresponds to the dataset radius, and ρ represents the maximal margin.

$$R(f) \leq R_{emp}(f) + \sqrt{c/n(R^2/\rho^2 - \log(1/\delta))} \tag{5.5}$$

In this context, the CNN and SVM used in the experiments here are assessed to understand the sample size they require to ensure learning in the context of speech recognition, allowing to estimate their expected risk value over unseen examples. The CNN architecture used is composed of one convolutional layer with 38 units whose mask size is 29×1, as estimated using the adapted FNN (FERREIRA *et al.*, 2018). The mask size and the number of units are important parameters to estimate the Shattering coefficient for a single CNN layer used in the experiments. Considering the formulation proposed in (Fernandes de Mello; Antonelli Ponti; Grossi Ferreira, 2018), Equation 5.6 defines the shattering coefficient for a single unit in the CNN layer, in which h is the space dimensionality and n corresponds to the sample size. Thus, Equation 5.7 corresponds to the Shattering coefficient for all 38 units in this layer, in which p is the number of hyperplanes.

$$f(n) = 2\sum_{i=0}^{h}\binom{n-1}{i} = 2\sum_{i=0}^{29\times1}\binom{n-1}{i} \tag{5.6}$$

$$CNN(n) = 2\sum_{i=0}^{h}\binom{n-1}{i}^{p} = 2\sum_{i=0}^{29\times1}\binom{n-1}{i}^{3}8 \tag{5.7}$$

Now, one can proceed by computing the generalization bound for the CNN (Equation 5.4), as shown in Equation 5.8. Considering $\delta = 0.05$, that represents a probability of 0.95 (i.e., 95%) to ensure that the empirical risk $R_{\text{emp}}(f)$ is a good estimator for the expected risk $R(f)$, meaning the error results measured for our classifier indeed work on unseen examples. Observe that the CNN requires at least $216,640$ examples to converge, while in practice, $1,447,869$ examples were available in training set.

In addition, another result from (VAPNIK, 2013) can be employed to prove that the CNN converges. Equation 5.9 considers the most relevant term to prove the learning convergence, analyzing Equation 5.8 (LUXBURG; SCHÖLKOPF, 2008; VAPNIK, 2013). Notice that as $\frac{\log CNN(n)}{n}$ approaches zero, term $\sqrt{4/n(\log(2CNN(n)) - \log(0.05))}$

from Equation 5.8 goes to zero, remaining the empirical risk as an assessment measure of the learning performance.

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{4/n(\log(2CNN(n)) - \log(0.05))} \tag{5.8}$$

$$\lim_{n\to\infty} \frac{\log\{CNN(n)\}}{n} \approx 0, \tag{5.9}$$

Next, the SVM is also analyzed considering Equation 5.5. In this case, we have an accuracy of 0.61 leading to $v(f) = 1 - \text{accuracy} = 1 - 0.61 = 0.39$, $R = 3,332,567$ (the radius estimated for the whole dataset), and $\rho = 173,869,050$ as the maximal margin found. Consequently, the generalization bound for our SVM is defined in Equation 5.10. Also, considering $\delta = 0.05$ as before, and $c = 4$ as taken in the default formulation (Equation 5.4), notice the SVM requires at least $1,476$ examples to converge, while in practice, $1,447,869$ examples were available in the training set.

$$\begin{aligned} R(f) &\leq 0.39 \\ &+ \sqrt{4/n(3,332,567^2/173,869,050^2 - \log(1/\delta))} \end{aligned} \tag{5.10}$$

Based on (Fernandes de Mello; Antonelli Ponti; Grossi Ferreira, 2018), the shattering coefficient for the CNN architecture that was composed of a single convolutional layer was formulated. The Shattering coefficient of the SVM was also calculated, according to SLT (LUXBURG; SCHÖLKOPF, 2008; VAPNIK, 2013). Those formulations ensure the framework presents learning guarantees in the context of speech recognition. No assumption is made about the joint probability because the training set was built from independent and balanced samples from a much larger corpus and care for is representation was taken; the labels assume deterministic values; the distribution of the joint probability is static since it is based on only this task and the distribution of the joint probability is unknown to the algorithm at the time of training.

5.1.3 Discussion

Most errors committed by the HTSVM were made between similar phonemes as found in (AMAMI; ELLOUZE, 2015), many of which unlikely would be perceived by human listeners. For example, the confusion between /s/ and /z/ are likely produced by the database itself where speakers may mix the two or use them interchangeably where it is semantically unimportant (HOFFMANN; TIK, 2009). This shows that the algorithm is quite promising since it avoided gross errors which are common in many modern acoustic modeling algorithms like GMM-HMM. The issue between similar phonemes seems

to be one which could be corrected by a robust language and/or pronunciation model. Also, since such a small window of 25ms was used, some of the vowels, especially diphthongs like /ay/ were damaged. Studies like (KARPAGAVALLI; CHANDRA, 2015) use 9 frames for classification and (YANG *et al.*, 2000) suggest at least 100ms for context which prompted (LOMBART; MIGUEL; LLEIDA, 2014) to use 11 frames (110ms) as context for classification. While our system is more accurate in frame classification than the system in (LOMBART; MIGUEL; LLEIDA, 2014), our PER is worse. This is probably also due to the small size where unnecessary insertions are made in the phoneme strings.

The combination of features extracted from single frames by a simple and shallow CNN classified by the HTSVM produced similar accuracy results to that study on a more difficult task as we were classifying full utterances which include pauses, co-articulation and a greater variation in pronunciation and not only single words and likely more careful speech. It is also important to note that some errors, like the case of the vowels which are close in the vowel space, could have been caused by the TIMIT transcriptions themselves where in some dialects these sounds could be very similar or even some sound in between. It was also interesting that the results are not far from the state-of-the-art forced alignment results where a pronunciation dictionary is employed for disambiguation. This shows that not only is the recognition robust but also that the CNN-HTSVM seems promising for segmentation as well either as a stand alone algorithm or a post-processor. We believe that the take home point here is that a quite robust acoustic model can be built even with a simple architecture and dataset when carefully constructed. By robust, it is meant that the model achieves 39.7% error in framewise classification and a 43.5% phone error rate using deep feature extraction and SVM classification even with little data (less than 7 hours). These results are comparable to studies which use well over ten times that amount of data. Beyond the intrinsic evaluation, the model also achieves an accuracy of 88% in the identification of epenthesis, the error which is most difficult Brazilian speakers of English, presenting a 69% relative percentage gain over the previous values in the literature.

5.2 Deep CNN-HTSVM with window-widening on TIMIT

The experiments in this section were designed specifically to establish a final architecture which would be useful within the scope of this project, first using TIMIT, since it is a good and small dataset which can be trusted and with the intention of expanding this method to more difficult databases like non-native or noisy speech. The architecture presented in Section 5.1 was modified in two areas: 1.) Inference capacity and 2.) Smaller dimension sizes. This was done with the intention of classifying larger windows of data. It was found, through the experimental phase that size alone does not better the capacity but it can reduce the feature maps with multiple layers and max-pooling, as presented in Table 15. This makes window widening a viable option, essentially.

Table 15 – CNN Dimension Improvements

Layers	Kernels	Ns	Max Pooling	Dims
1	29×01	38	5×5	988
3	15×02	36		**488**
	15×01	31	3×3 / 2×2	
	08×01	15	3×3 / 2×2	

Since a central point of this work is to show how careful parameter selection is useful for meaningful feature extraction from small datasets, the experiments, presented in Table 16 were designed to show this. They were designed to address the concept that deeper networks generalize better (YU *et al.*, 2012). The parameters in the network architecture, which yielded both networks from Table 15, were estimated using an approach based on False Nearest Neighbors (FNN) proposed by (FERREIRA *et al.*, 2017). This technique estimates the kernel sizes to best represent data patterns. Then the best fitting network which properly passes information from one layer to the next was built, specifically that which has the least number of units and large kernel sizes, and finally, found the most aggressive pooling layers to lower the dimensions of our feature maps since these would then be passed on to a SVM in multiple frames, which provides the supervised learning guarantees needed (VAPNIK, 1998). ReLU was applied as the activation function for both versions due to its widespread adoption in literature.

Table 16 – Deep CNN architectures

#Layers	Units	FA	F1
1	38	**0.430**	**0.225**
2	38-38	0.375	0.199
3	38-38-38	0.312	0.134
5	38-38-38-38-38	0.216	0.056
1	256	0.403	0.245
2	256-38	0.376	0.198
3	1024-512-256-112-38	0.129	0.075

The shallow network took less than half an hour to execute on a computer with 8GB of RAM and conventional hardware (Intel i7), whereas the deep network required 4 hours using a single Titan-X GPU and 32GB of RAM. The same structure which was successful in the experiments from Section 5.1 was tested for one layer with different layer configurations 1, 2, 3, and 5. Then, some typical unit configurations were used for CNN with bottlenecking into our minimal layer. This architecture was derived from anecdotal experience with fellow researchers and a synthesis of some of the popular work in the area like (HINTON *et al.*, 2012), (SAINATH *et al.*, 2013) and (ABDEL-HAMID *et al.*,

2014). All CNN features presented here were trained using the same fully-connected MLP employed in Shulby *et al.* (2017) for easy comparison. The decision was made not to compare regularization techniques, such as dropout, since they force a restriction on the hypothesis space, which does can only provide approximate solutions. The purpose of these experiments was simply to show how augmented architecture affects model capacity and saturates the information passed between layers.

The best classifier for the CNN features was a 4^{th} order polynomial kernel with $coef0 = 1$ (as a non-homogeneous kernel) and a cost $C = 10,000$. This was also left unchanged from Section 5.1. Since the SVM is less agile for training multiple experiments, we used a MLP, also the same as in Section 5.1, to find the most cost effective window sizes and compared those results with some chosen SVM classifiers. We also added some variations in our experiments like a stochastic gradient descent (SGD) (the only difference from the previous MLP) solver as a regularizer for the MLP classifier and skipped some frames, similar to the approach by (LOPES; PERDIGÃO *et al.*, 2009). The difference in the current frame skipping technique is the two adjacent frames (1 right, 1 left) were left always intact, believing that these frames are the most important ones. After the three middle frames, every 2 frames were skipped until the extremity of the window was reached.

5.2.1 Results

Experiments with window skipping are noted in Table 17 by an asterisk symbol in the number of frames column. It should also be noted that a SGD solver was also applied when training the CNN which is why the single frame MLP appears with slightly better results here than in subsection 5.1.

Table 17 – Experiments on CNN features with window-widening

No. Frames	Classifier	FA	F1
1	MLP	0.51	0.36
3	MLP	0.55	0.41
5	MLP	0.57	0.45
7	MLP	0.59	0.48
9	MLP	0.60	0.49
11	MLP	0.61	0.51
11*	MLP	0.61	0.50
11*	MLP w/SGD	0.62	0.51
13	MLP	0.61	0.51
15*	MLP w/SGD	0.63	0.52
19	MLP w/SGD	0.63	0.52
1	SVM	0.58	0.44
3	SVM	0.59	0.48
11*	SVM	**0.70**	**0.61**

5.2.2 Discussion

With simply one frame for classification, the CNN features were classified similarly to the studies like (LOMBART; MIGUEL; LLEIDA, 2014), who use 11 frames (110ms) and (LOPES; PERDIGÃO *et al.*, 2009) who use as many as 17 frames as context for classification. As the number of frames was increased, superior results were produced. It is my belief that 11 frames seems to be the most cost-effective number since larger configurations demand a much larger number of feature maps and offer little improvement. The result of 70% is also comparable to the state-of-the-art approach proposed by Graves and Schmidhuber (2005) at 70.2%, which is still considered the best results on TIMIT. In addition, the current method guarantees generalization which Graves and Schmidhuber (2005) does not do. The multiple max-pooling layers were great facilitators for the window widening technique since without this dimension reduction the SVM would have been prohibitive to train and even the MLP would have been very difficult. Therefore, while more layers do not seem necessary for better results, they can help reduce dimensionality. We also show that larger architectures with out of the box configurations are harmful to model capacity and that when dealing with limited data, careful parameter selection and attention to supervised learning guarantees are essential for building robust models.

5.3 Final CNN-HTSVM for Non-native Speech

After a series of experiments and promising results on the TIMIT dataset, the pipeline was finally tested on the dataset developed for this dissertation, specifically,

non-native speech from Brazilian learners of English in their own (sometimes noisy) environments. As mentioned earlier, two research questions must be answered:

1. Does the model achieve results which are better than the SotA and bring us closer to the results produced by manual alignment?; and

2. Is the model robust?

For the first question we look towards intrinsic results and compare them with the SotA experiments for both native and non-native speech in Section 5.3.1. Here we can compare the results from the validation set of the Ramble corpus to its manually segmented key and make comparisons to the work in the literature about native and non-native benchmarks. For the second question, extrinsic results for a specific example will be shown. This is done to show how the model would perform on a specific task, in particular, on a well known L2 error made by Brazilian learners of English. Thorough analysis of this error shows just how robust speech recognition could be applied to a pronunciation tutor. The error selected is the one which creates the most confusion for listeners, namely epenthesis, where syllables are simplified, often with the infamous addition of the /i/ phoneme in the word initial or syllable final positions.

5.3.1 Intrinsic Results

Intrinsic evaluation is the most common type of evaluation for most machine learning algorithms. This is in part due to the fact that it is relatively simple to carry out. In the studies presented in Table 18, a training set is used to train each model and a validation set is used to intrinsically evaluate it. With the exception of this dissertation and the one by Garber, Singer and Ward (2017), the TIMIT benchmark is used. This is a good benchmark for speech recognition but unfortunately it does not address the issue at hand in the other studies in this table. Non-native speech has yet to establish a strong benchmark. This can be difficult because so far, most approaches to non-native speech are language dependent which requires a test set from a similar population. Still, we can try to draw some conclusions based on the intrinsic evaluations of the Ramble corpus and the results mentioned earlier in this thesis, using the same method on the TIMIT corpus.

Table 18 – Comparison of the current method with SotA Results using PER and FER on native (N) and non-native(NN) speech

Name of Study	**Method**	**Size**	**N/NN**	**PER**	**FER**	**F1**
Graves, et al., 2013	BLSTM-RNN w/PT	4h+PT	N	17.7	27.88	-
Song & Cai, 2015	CNN + CTC	4h+PT	N	29.4	22.1	-
van Niedek, et al., 2016	DLSTM-RNN Replication	4h+TFW	N	25.4	29.4	-
Shulby, et al., 2018	CNN-HTSVM	4h	N	33.0	30.0	0.61
Garber, et al., 2017	HMM-GMM	95h	NN	39.2	-	-
Ramble (this dissertation)	CNN-HTSVM	7h	NN	43.5	39.7	0.50

From the results in Table 18, we can see that the results are of similar proportion for non-native speech as with native speech when compared with the results of the same method in (SHULBY *et al.*, 2018 (Submitted)). This is probably due to the larger number of classes in the non-native corpus. Also many sounds are very difficult to separate in Brazilian speakers, particularly nasalized vowels and rhotics. This also creates a very dynamic vowel space since some realizations are simply something in-between which is difficult to label correctly. It is also interesting that the results are very similar in metrics to the results obtained by Garber, Singer and Ward (2017), considering that the current scenario is more difficult. It should be noted that the tasks in this study and the Garber study are not similar at all. Here, the model was trained on an unrestricted speech domain, whereas the Garber study only uses air-traffic control data, making their task much easier. Here we have much less than a tenth of the data, more noise, and we do not specify a domain. This is promising and shows that with some work, these results could likely be improved. As was done in the previous experiment, the confusion percentages are shown in Table 19.

Table 19 – Most Frequent FER Confusion percentages

CNN-HTSVM		
True Phoneme	Predicted Phoneme	Confusion Rate
ah	ow	13.08%
iy	ih	12.85%
ih	ay	13.49%
eh	ay	18.68%
iy	ay	11.92%
ah	eh	10.76%
d	t	12.67%
ey	ay	25.7%
t	s	8.86%
ah	ih	9.71%
o	aa	32.6%
er	ah	17.13%
iy	w	9.05%
k	t	9.96%
eh	ae	11.43%
ah	ihm	5.93%
er	aa	10.38%
ah	aa	5.64%
iy	m	5.99%
ahm	ihm	18.04%

The results from Table 19 confirm the suspicion that the vowel space was problematic. Most errors occurred on similar sounding vowels and vowels which appear orthographically in unexpected situations for Brazilians. This is probably the cause of “e” and “a” vowel which were often confused for the phoneme /ay/, also one can see some nasal errors in the list. Even though these phonemes were less frequent that were responsible for a large number of errors. The manual segmentation could partially be at fault here since it is rather confusing, even for a specialist, as to where a vowel ends and a nasal begins in Brazilian speech. The phoneme /ah/ is also confused for several vowels. This is probably because the vowel does not exist in Brazilian Portuguese so it is included in unexpected places. The pronunciation dictionary augmentation algorithm did not have a rule for this case so maybe that is a point of improvement. Some stops were surprisingly confused as was the case in the TIMIT corpus. This type of error seems to be easily dealt with in later modules and could have to do with some noise. Still, it seems that the take home point in the table is that further improvements for vowels would greatly boost the accuracy of the model.

5.3.1.1 Hyperplane Separation and Convergence analysis on the RAMBLE data

The current algorithm was proven to converge even with a small number of samples on the TIMIT corpus in 5.1.2. Here we will go into more detail of the hyperplane separation within the hypothesis space and rerun the convergence analysis for the RAMBLE data. First we will investigate the separation of each level of our SVM tree as shown in the Classification Section in Chapter 4. In order to simplify the representation, the Principal Components Analysis (PCA) algorithm was used to identify the most discriminant CNN extracted features so that the hyperplanes could be plotted on a two dimensional space. Also we used 10,000 samples to train and 2,000 samples for testing as we used in our pilot experiments for each classifier. This was done for two reasons: 1. to speed up the experiments; 2. to show how the SVM is able to generalize the data well even with few examples. For comparison, the same process was used for the MLP classifier used in the TIMIT experiments to show how the SVM decision boundary provided for a robust separation.

Level 1

The first level consists of the separation of obstruents, sonorants and silence. Even though silence was not used in the final calculation of FER and PER, it is useful in the industry for practical applications like phonetic boundary segmentation and Voice Activity Detection (VAD).

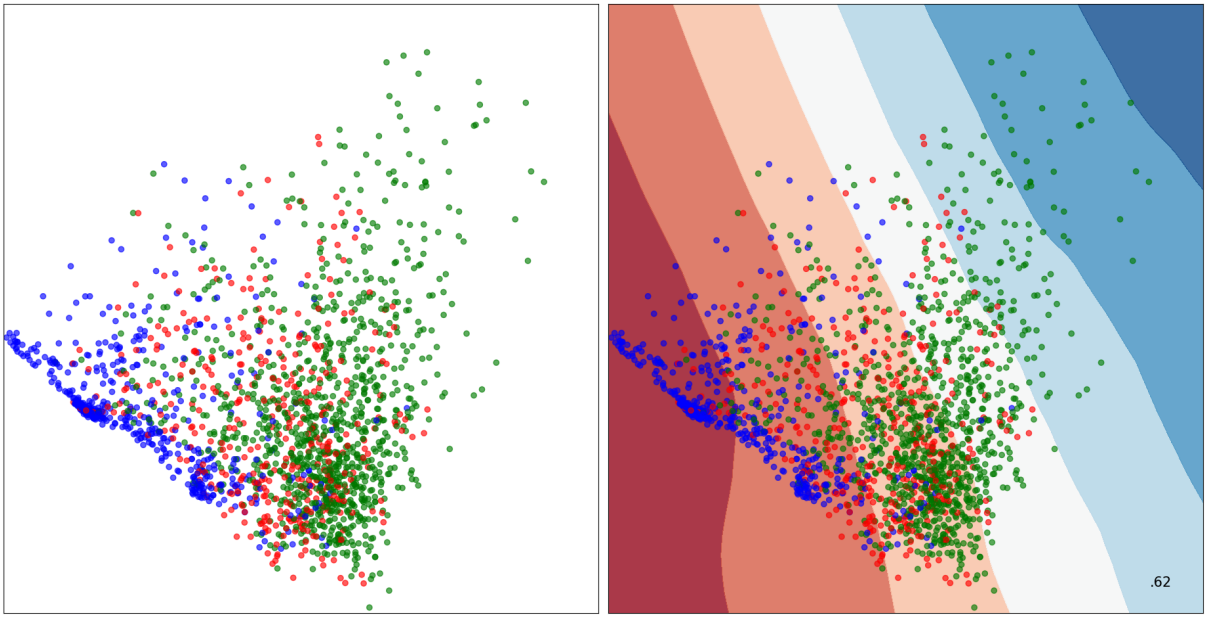

Figure 26 – MLP Obstruents, Sonorants and Silence

This first classification is a bit more straight forward than others and with more data, the classification becomes quite clear in the final model with all of the training data (about 97%). Still, we can see that the SVM generalizes a little bit better.

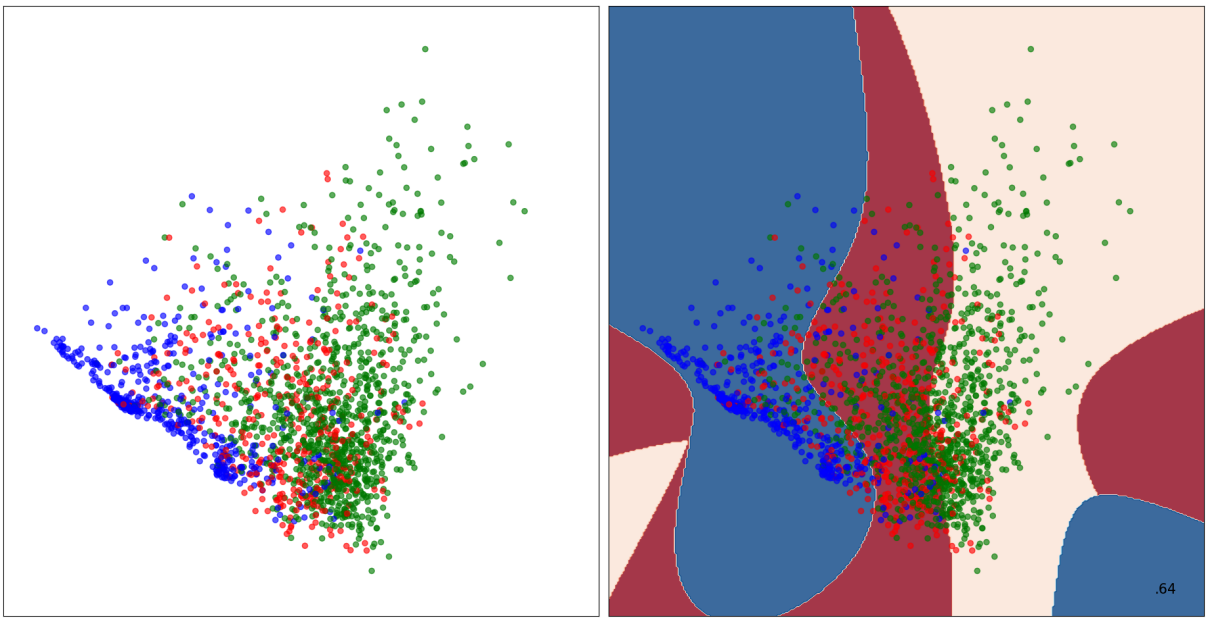

Figure 27 – SVM Obstruents, Sonorants and Silence

Level 2

On the second level, broad phonemic types are defined in the groups: 1. /hh/; 2. fricatives; 3. stops; 4. affricates, all from the obstruents branch; then from the sonorants branch 5. liquids and glides; 6. nasals and finally; 7. vowels. It is important to explain these divisions since there are multiple different ways this could be done using the same knowledge-based arguments. Perhaps, the biggest question is why the /hh/ phoneme is already being classified as the first and only phoneme classified up until now.

It was found in experiments that the /hh/ phoneme was confused at times with other aspirated phonemes like /p/ and /k/ and even some noisier fricatives like a harsh /f/ sound. It is also in a class of its own as a glottal fricative and considered much different in articulation than the other fricatives. So the main decision was whether to classify it in level two or to send it along as a fricative and classify it in level three. From a linguistic point of view the latter still sounds more interesting, even though some debate may exist, but from a machine learning point of view, one can have a strong preference for early classification. The first motivation is that it is one less class to deal with in a second classifier. This already avoids the chance of error from a second classification. Also, it minimizes the number of samples which go on for training, the more classes can be broken up easily, the simpler it will be as the algorithm goes on. The difference between fricatives and stops is quite fundamental, so it is easy to justify this separation. Affricates could certainly be grouped together with fricatives but since they already posses some fundamentally discriminative characteristics, it also make sense to make this distinction earlier rather than latter.

The results here were a bit surprising since the final results are so high. The final

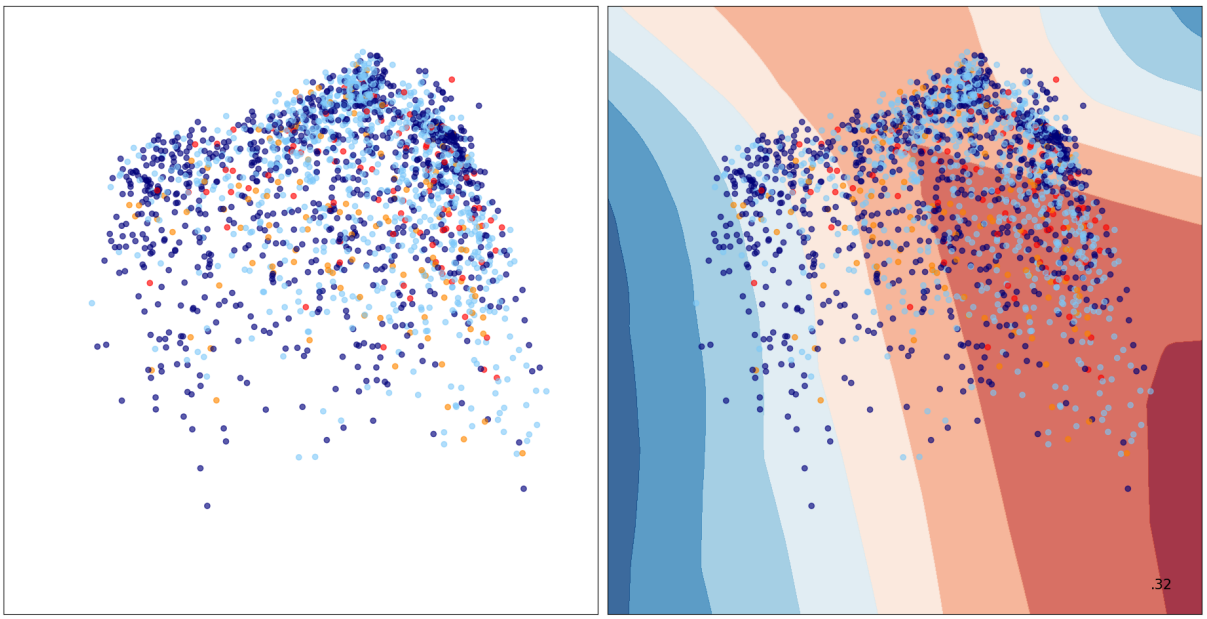

Figure 28 – MLP /hh/, Fricatives, Affricates and Stops

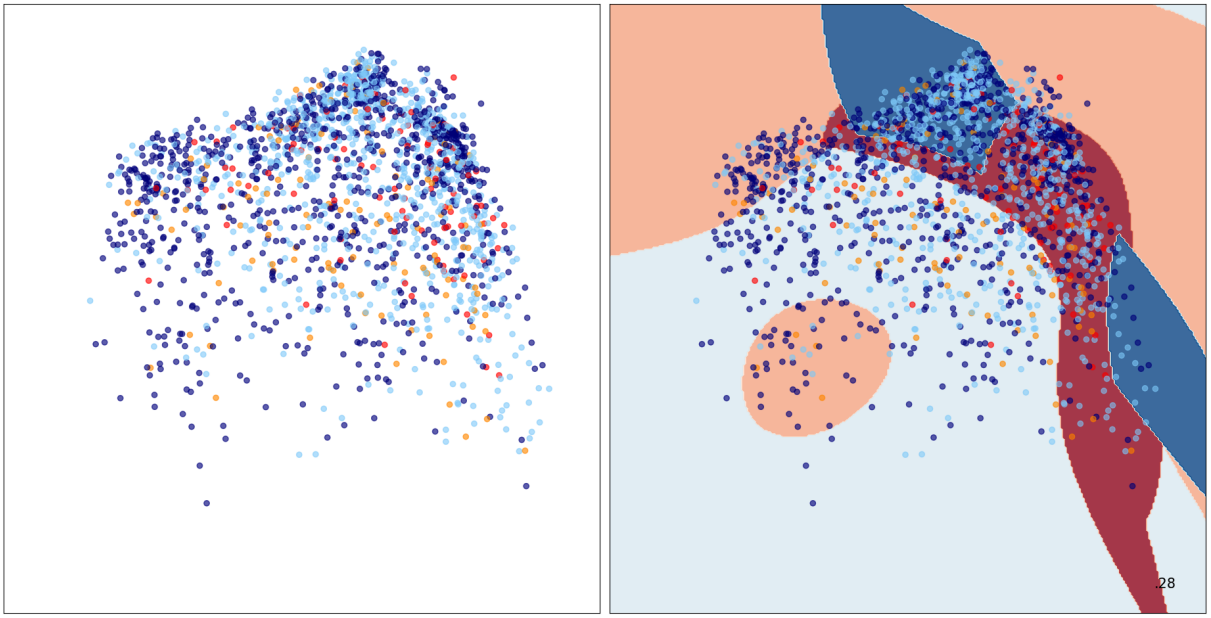

Figure 29 – SVM /hh/, Fricatives, Affricates and Stops

MLP model achieved an accuracy of 85% and the SVM 96%. It seems that his problem needs more dimensions to separate well. This explains why the MLP seems to do better here, but if one observes closely, the gradient is actually quite different from the decision boundaries and upon closer inspection, it seems that the larger class “fricatives” was heavily preferred by the model, where this was not the case in the SVM.

The case of nasals, liquids/glides and vowels is a similar story, where the MLP preferred “vowels” in over half of the classifications with this dimensionality and was unable to separate the other classes. The SVM was also not successful. The final SVM

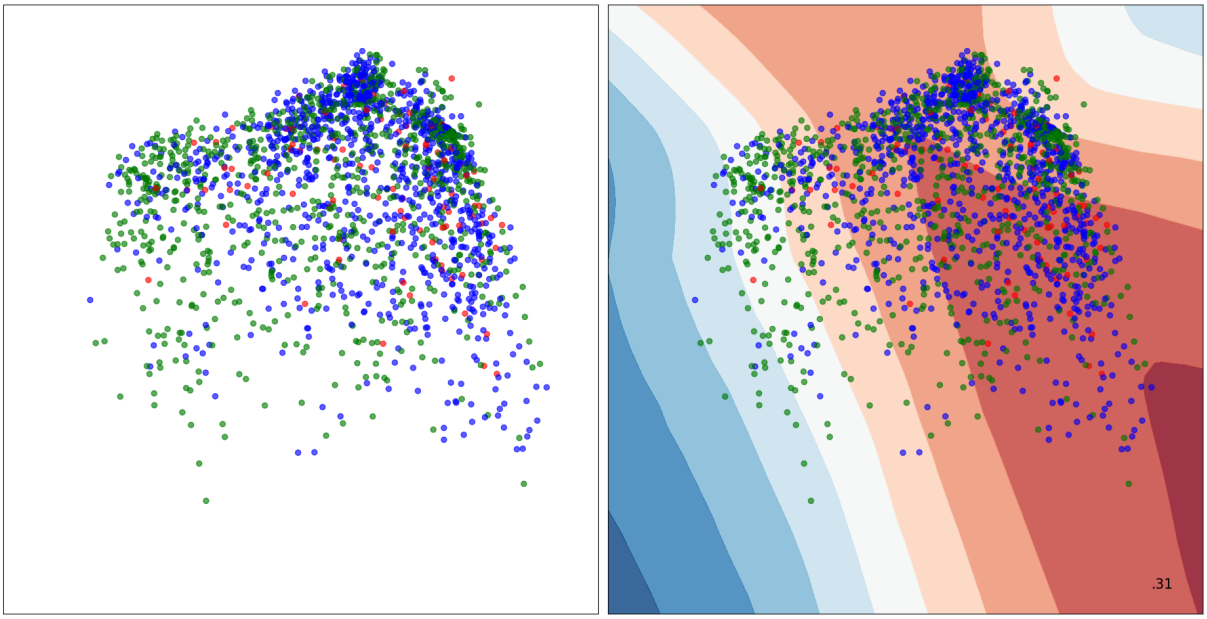

Figure 30 – MLP Nasals, Liquids/Glides and Vowels

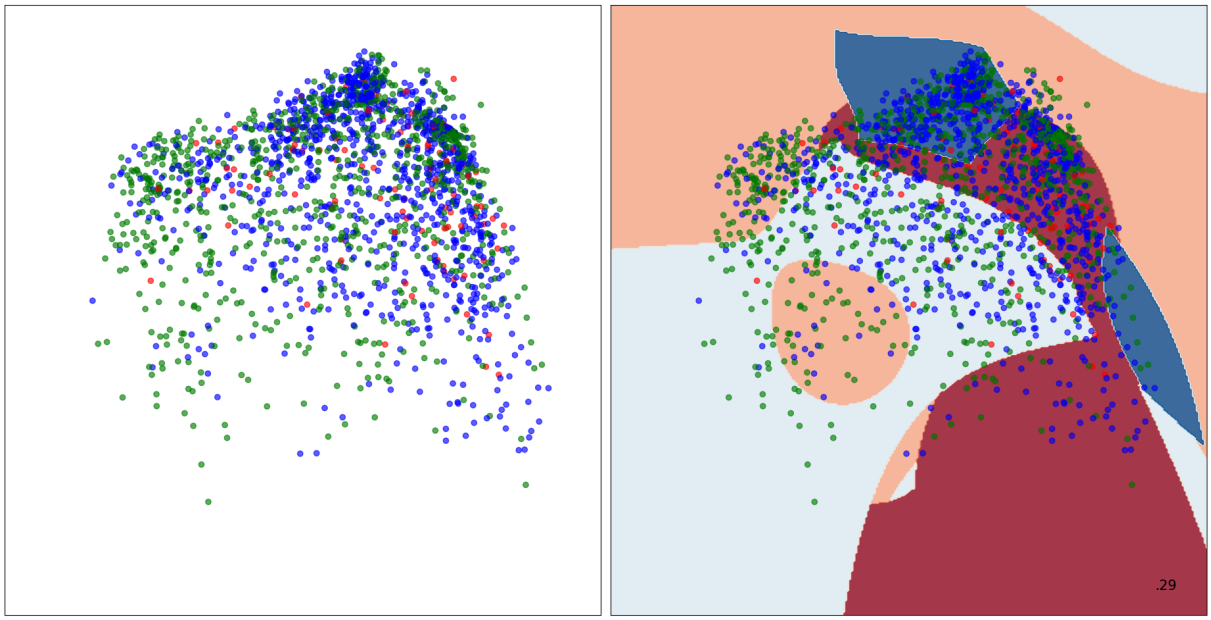

Figure 31 – SVM Nasals, Liquids/Glides and Vowels

model achieves 93% in accuracy and the MLP model ends up 10 points behind at 83%.

Level 3

The third level is where a great deal of classification really occurs. Experiments were run to try and divide these classes further in groups like voiced/unvoiced but two classifications only damaged results so some larger groups were simply classified at this stage.

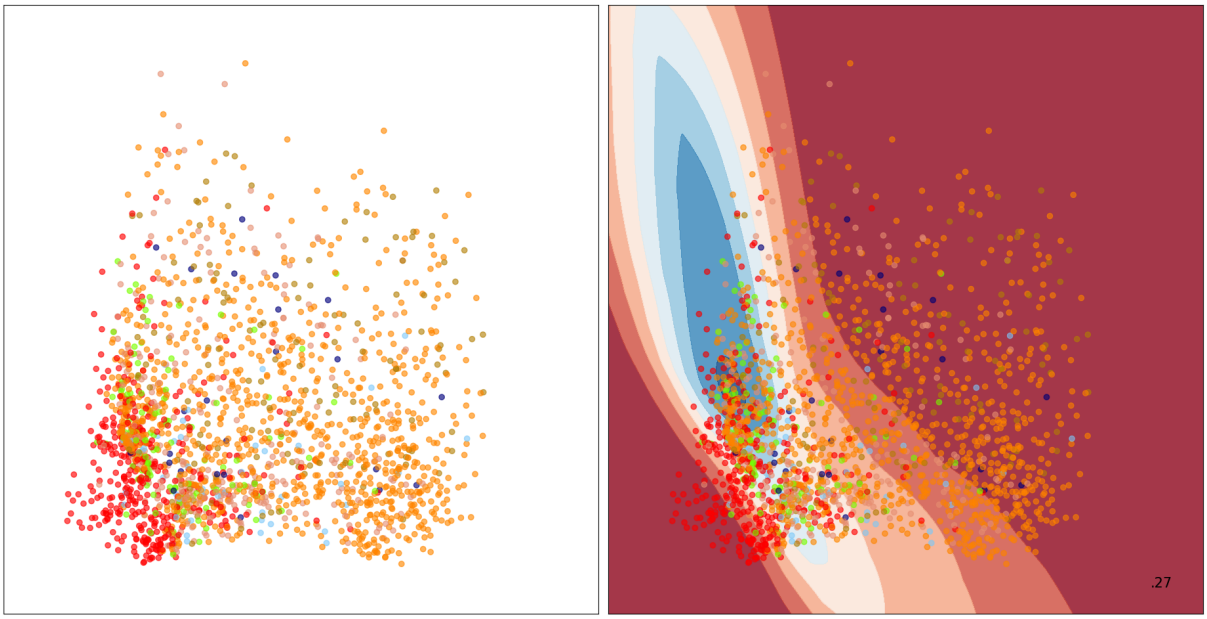

Figure 32 – MLP Fricatives - /zh/, /dh/, /z/, /f/, /s/, /sh/, /th/ and /v/

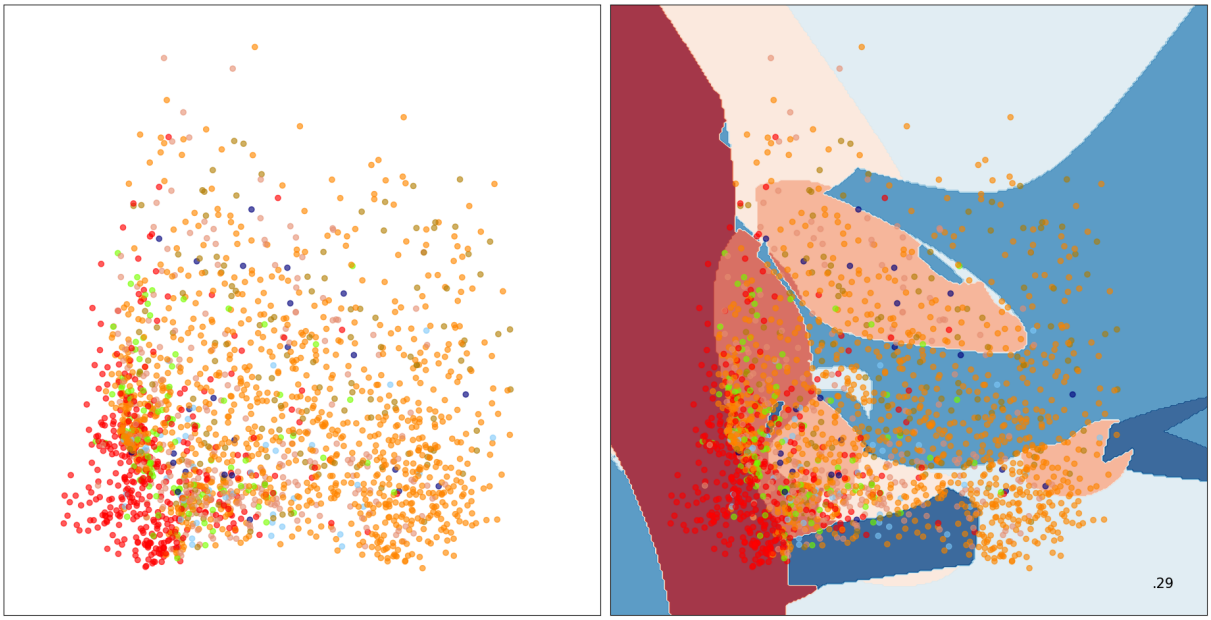

Figure 33 – SVM Fricatives - /zh/, /dh/, /z/, /f/, /s/, /sh/, /th/ and /v/

The fricatives have a large number of classes but actually does a good job in the final model where the SVM was able to correctly identify 75% of the frames, where the MLP had more difficulty at 63%.

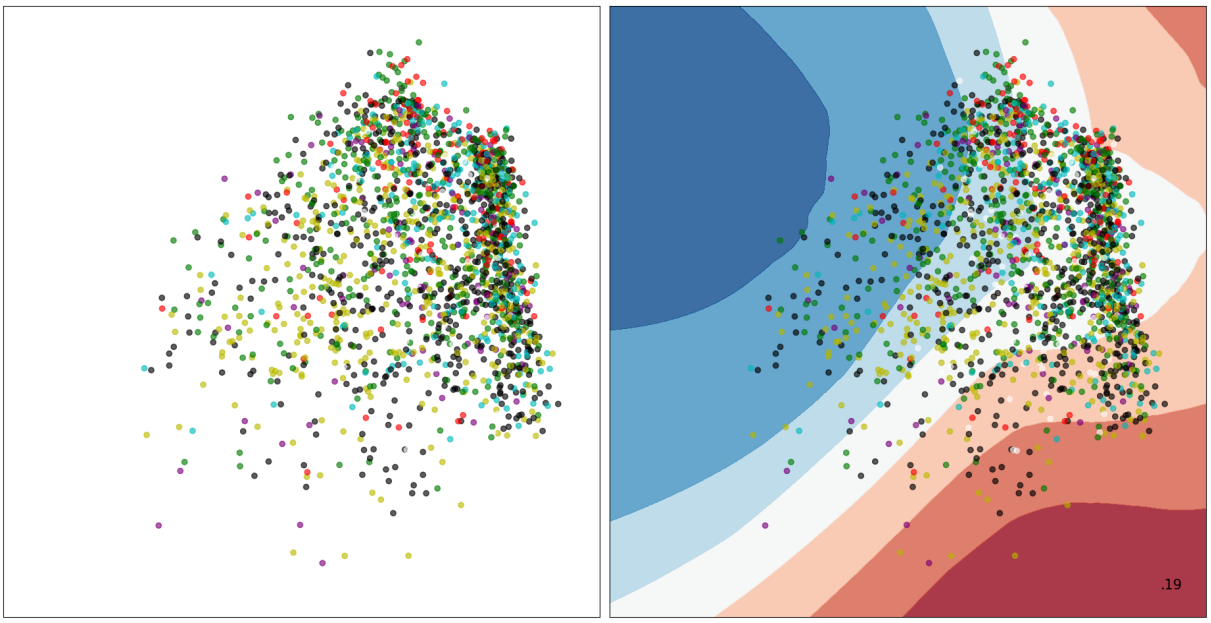

Figure 34 – MLP Stops - /b/, /d/, /g/, /k/, /p/, /t/ and /tt/

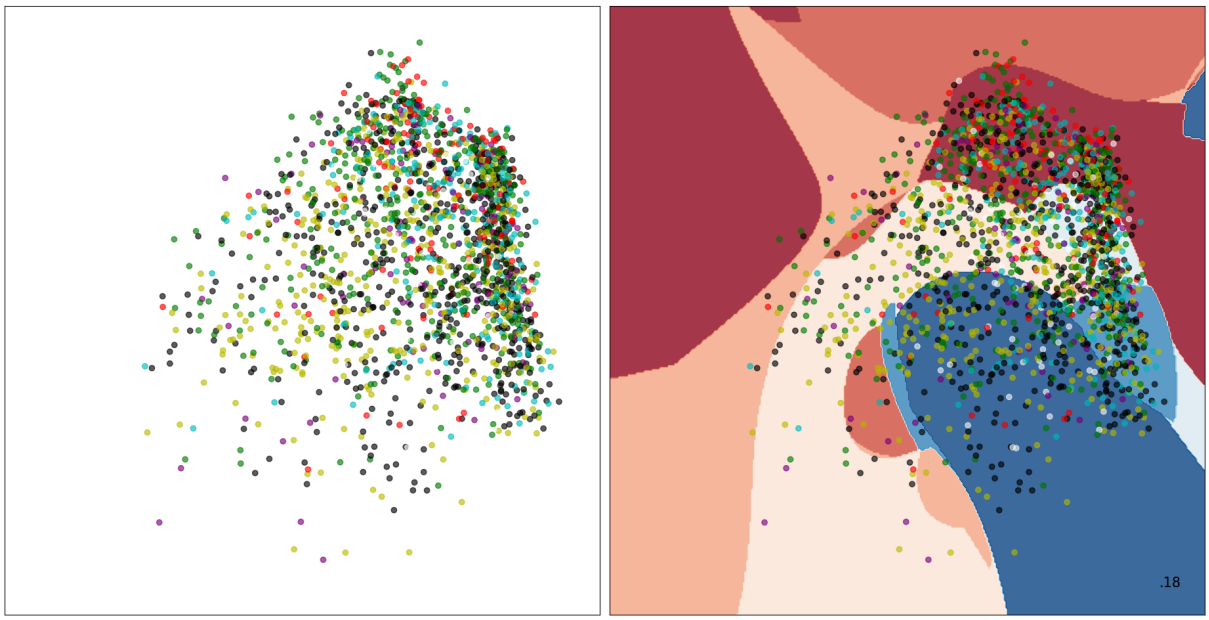

Figure 35 – SVM Stops - /b/, /d/, /g/, /k/, /p/, /t/ and /tt/

The case of stops is actually a difficult case. This makes sense since it is also not easy for a linguist either. The identification of the existence of a stop is rather simple since the explosion is very clear in the waveform but most linguists use information from the surrounding sounds to classify the actual type of stop. It would be interesting to experiment with larger window sizes or even multi-modal data. It is also the only final model where the MLP out performed the SVM. The MLP was able to get 71% of the frames correct where the SVM came in a bit shorter at 69%.

One can see here that affricates in fact were rather easy. It made sense to separate

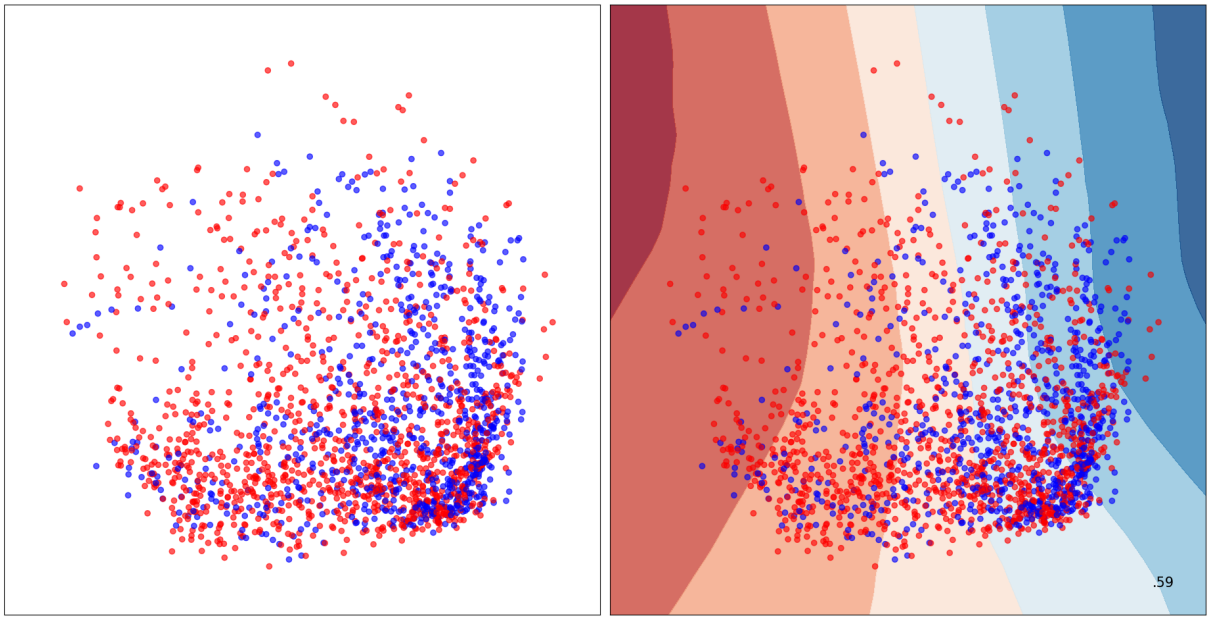

Figure 36 – MLP Affricates - /ch/ and /jh/

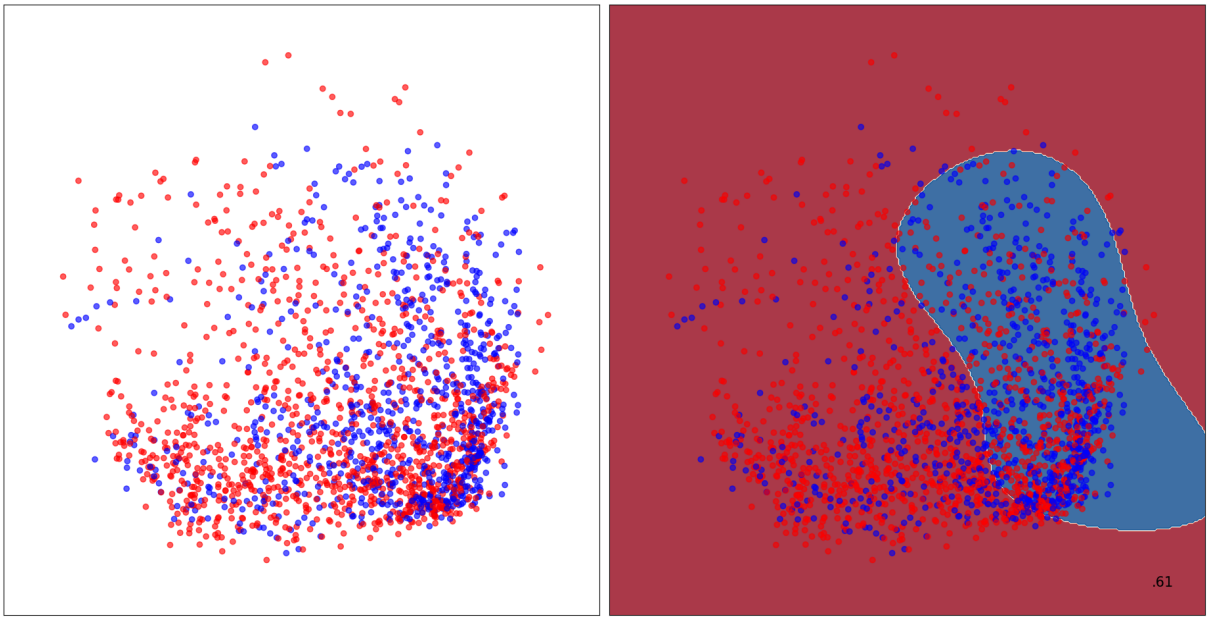

Figure 37 – SVM Affricates - /ch/ and /jh/

them from fricatives which already had a large number of classes and to classify them with a high level of accuracy. The final SVM and MLP models achieved 91% and 90% frame accuracy respectively.

For the sonorants we have only three categories. Liquids and glides are grouped together since they are quite different from other sonorants, especially in English where they possess more consonant-like characteristics than in Portuguese. Also the less classes to go on as vowels, the better the generalization is likely to be. Finally, vowels share that they are the typical sonorant (pitch is well represented) and, other than some Portuguese

glides and nasals, can be easily set apart.

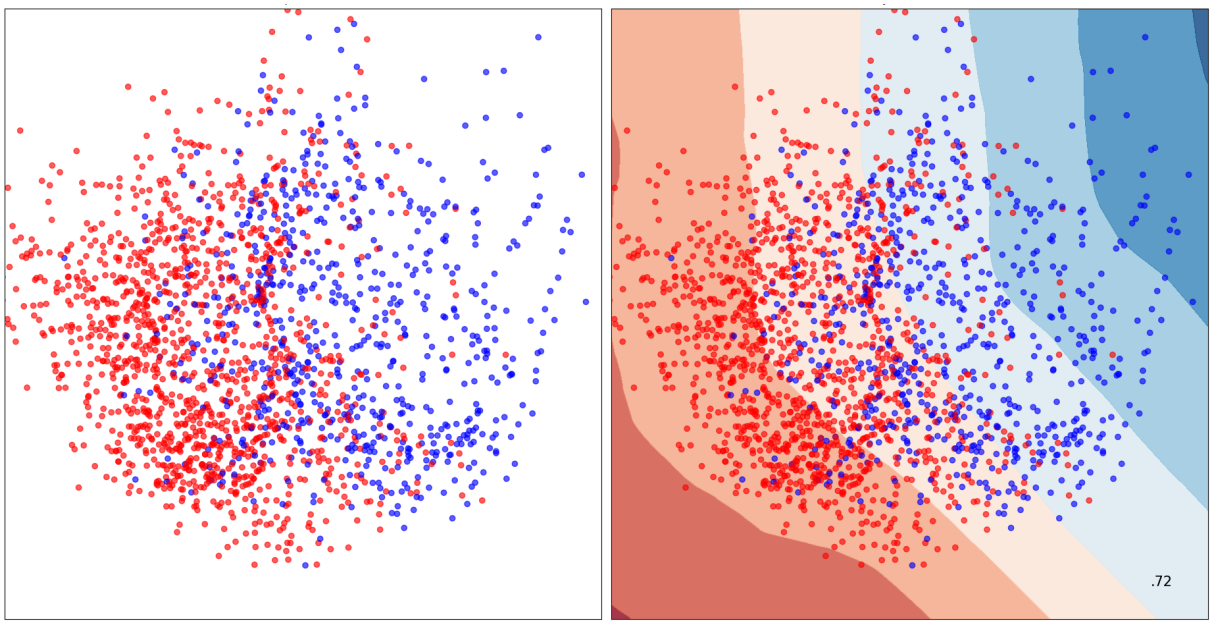

Figure 38 – MLP Liquids and Glides

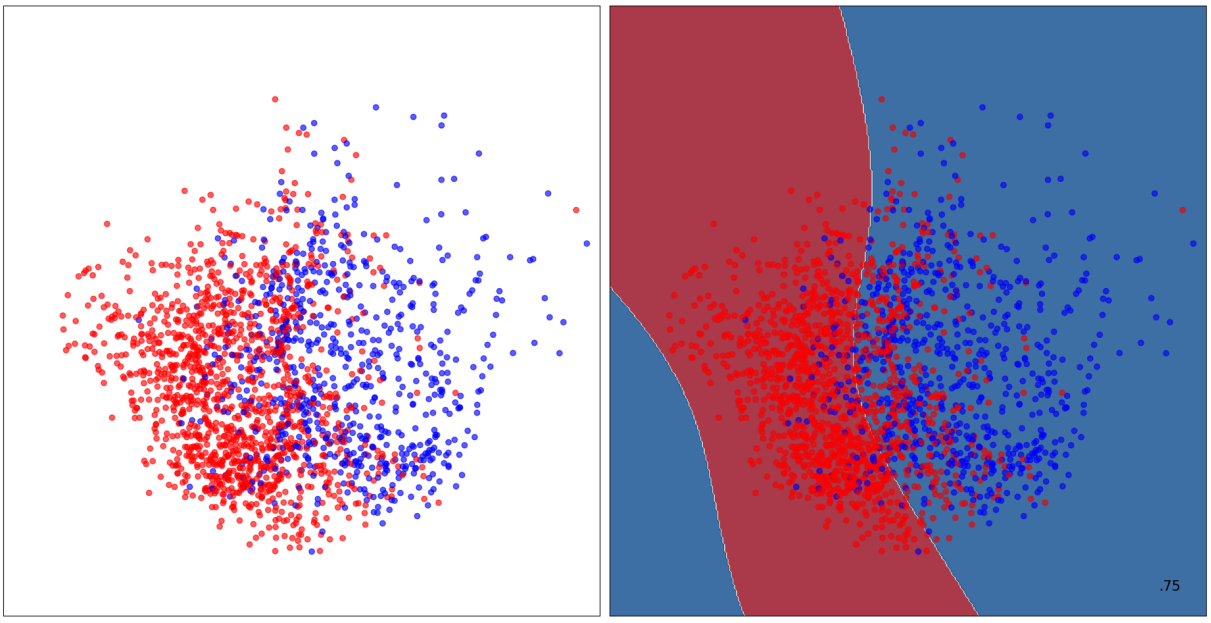

Figure 39 – SVM Liquids and Glides

Liquids and glides were another easy case. Here, even with 10,000 samples and only two dimensions the classification is already quite positive. The final SVM and MLP models achieved 97% and 96% respectively.

In this dimensionality the nasals seem more difficult to classify. These sounds are also not so easy for a human but with enough context near neighboring sounds, it is certainly easier than the case of stops. In this model the SVM outperforms the MLP in the final model at 88% to the MLP's 72%.

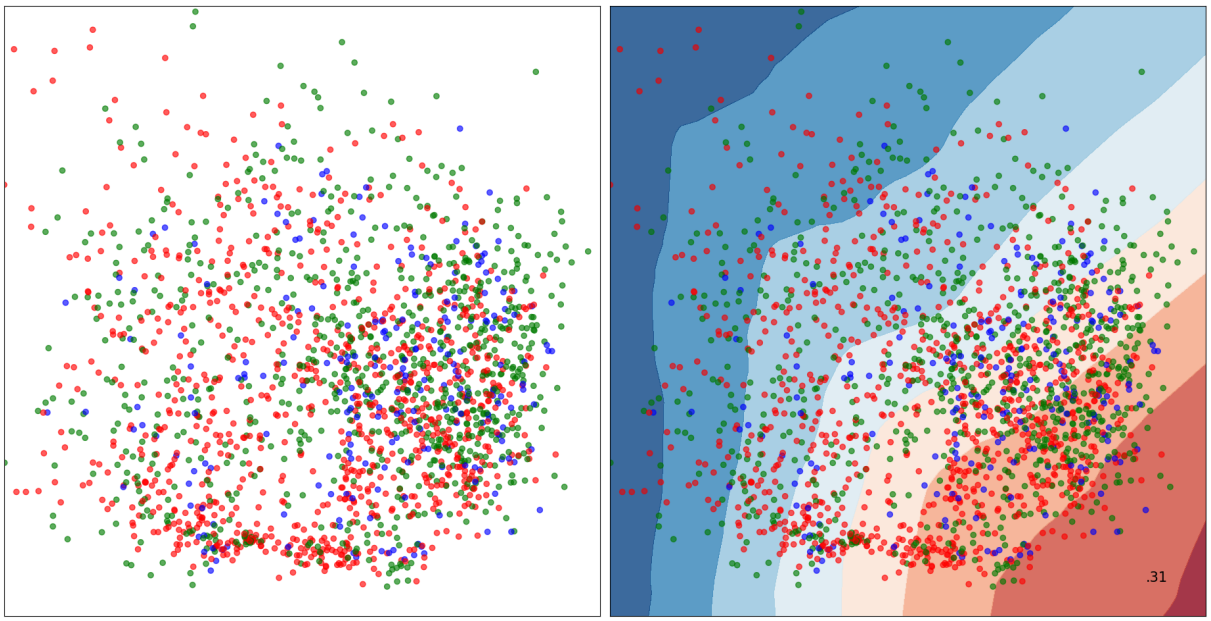

Figure 40 – MLP Nasals

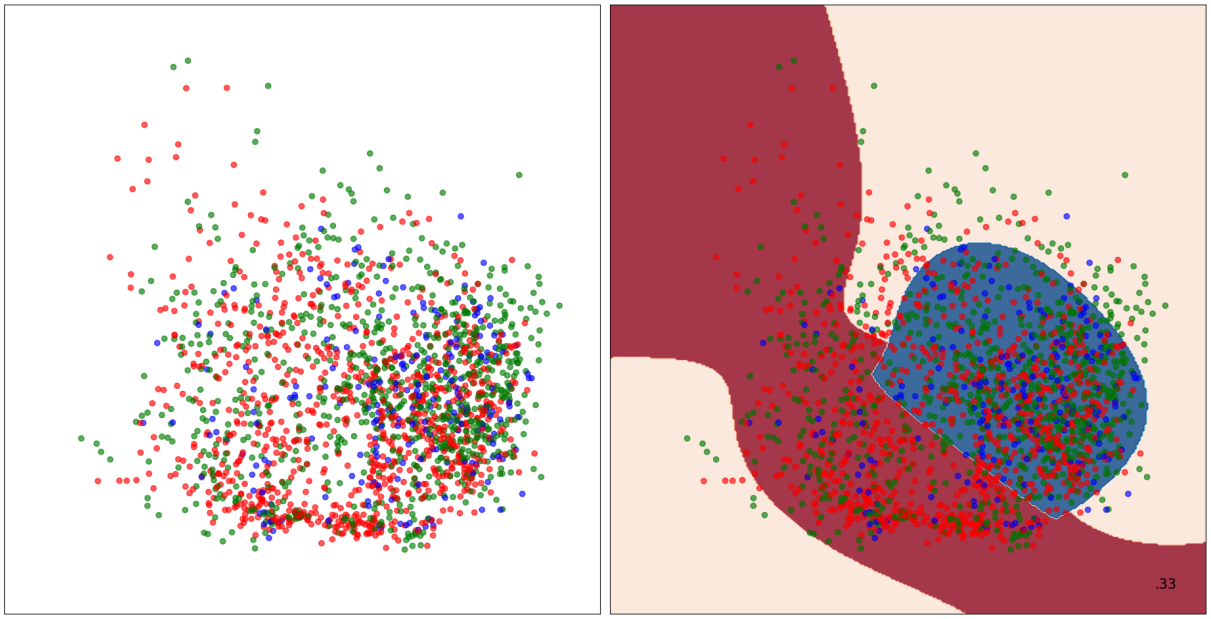

Figure 41 – SVM Nasals

Finally, we also classify the vowels. On the TIMIT dataset, the vowels were divided as front, central and back vowels. In the RAMBLE corpus, this division became more complicated with the large variety of semi-vowels and nasalizations which are carried over from Portuguese. This is the problem child of the corpus and one which surely would benefit from continued study.

Here we can see how chaotic the vowel space is. Granted the low-dimensionality leaves a great deal of valuable information out but even the final models struggled. The SVM was able to classify 46% of the frames correctly, where the MLP was only able to

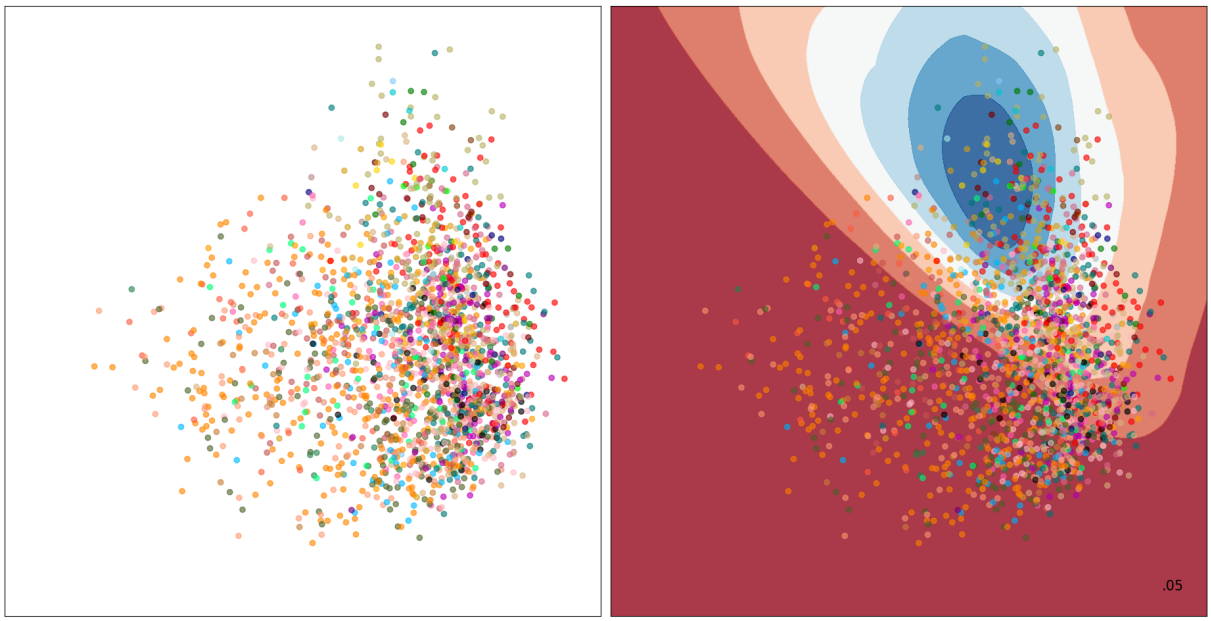

Figure 42 – MLP Vowels - /aa/, /aam/, /aar/, /ae/, /aem/, /ah/, /ahm/, /ao/, /aw/, /awm/, /ay/, /aym/, /eh/, /ehm/, /er/, /ey/, /eym/, /i/, /ih/, /ihm/, /im/, /iy/, /iym/, /o/, /om/, /or/, /orm/, /ow/, /owm/, /oy/, /uh/, /uw/ and /uwm/

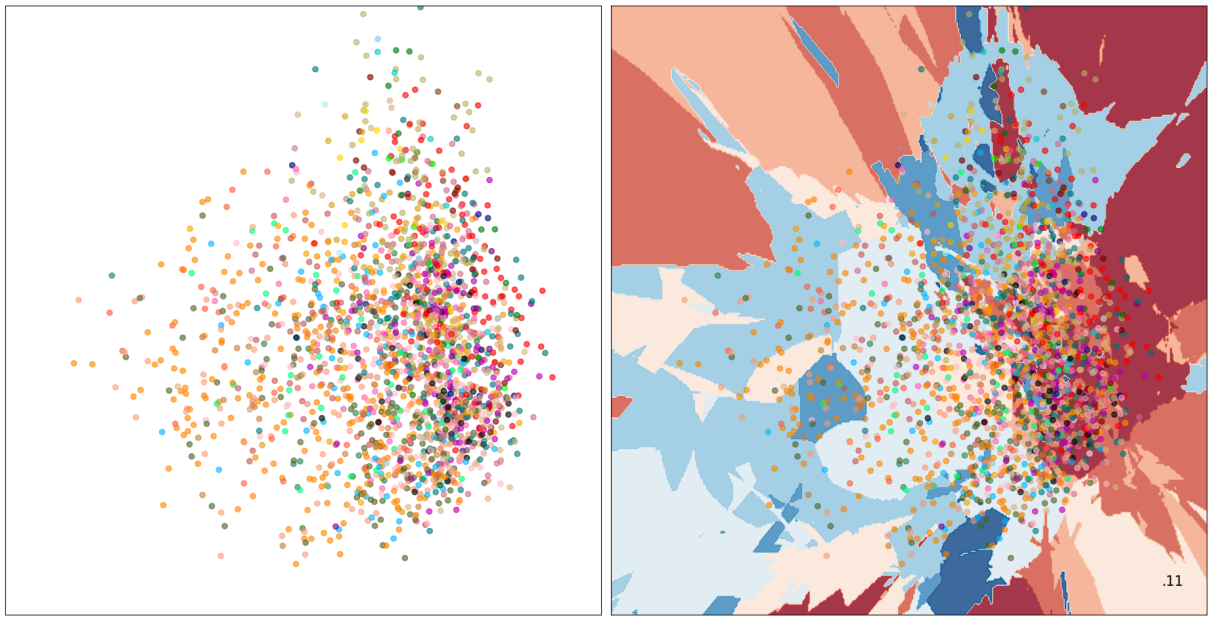

Figure 43 – SVM Vowels - /aa/, /aam/, /aar/, /ae/, /aem/, /ah/, /ahm/, /ao/, /aw/, /awm/, /ay/, /aym/, /eh/, /ehm/, /er/, /ey/, /eym/, /i/, /ih/, /ihm/, /im/, /iy/, /iym/, /o/, /om/, /or/, /orm/, /ow/, /owm/, /oy/, /uh/, /uw/ and /uwm/

get 35% correct. Some of this is handled well when the PER is calculated and possibly are less of an issue when a pronunciation model is used. Still, it is clear that there is room for improvement here.

Level 4

Level four simply finishes off the classification of liquids and glides since the vowels were taken care of in level three.

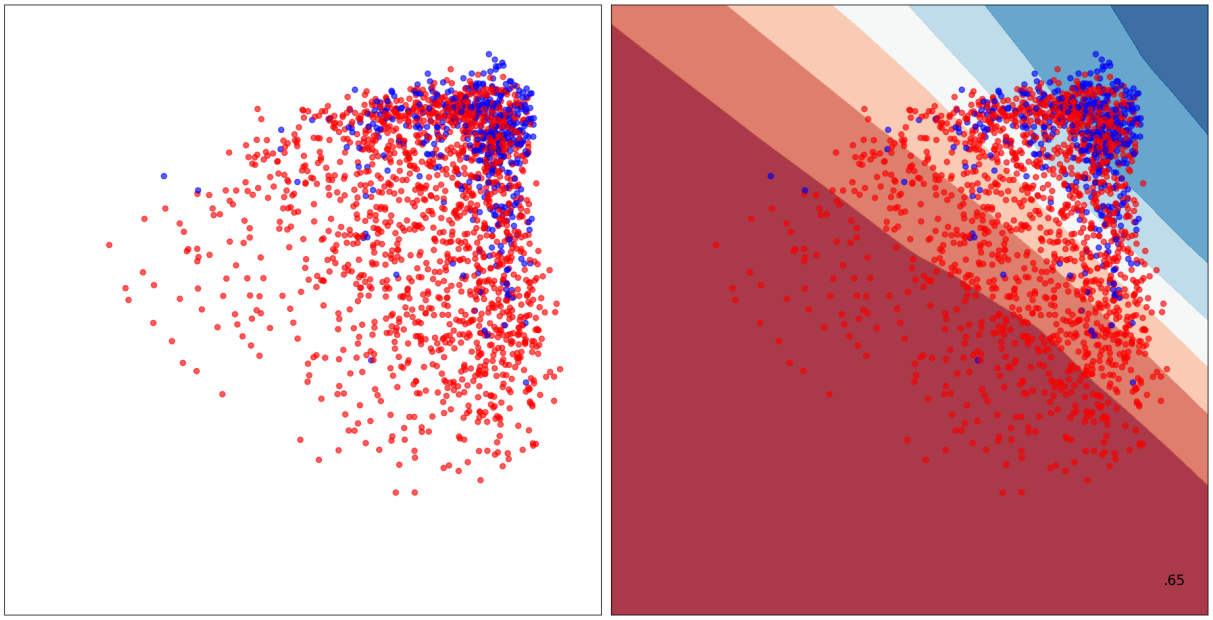

Figure 44 – MLP Liquids - /l/ and /r/

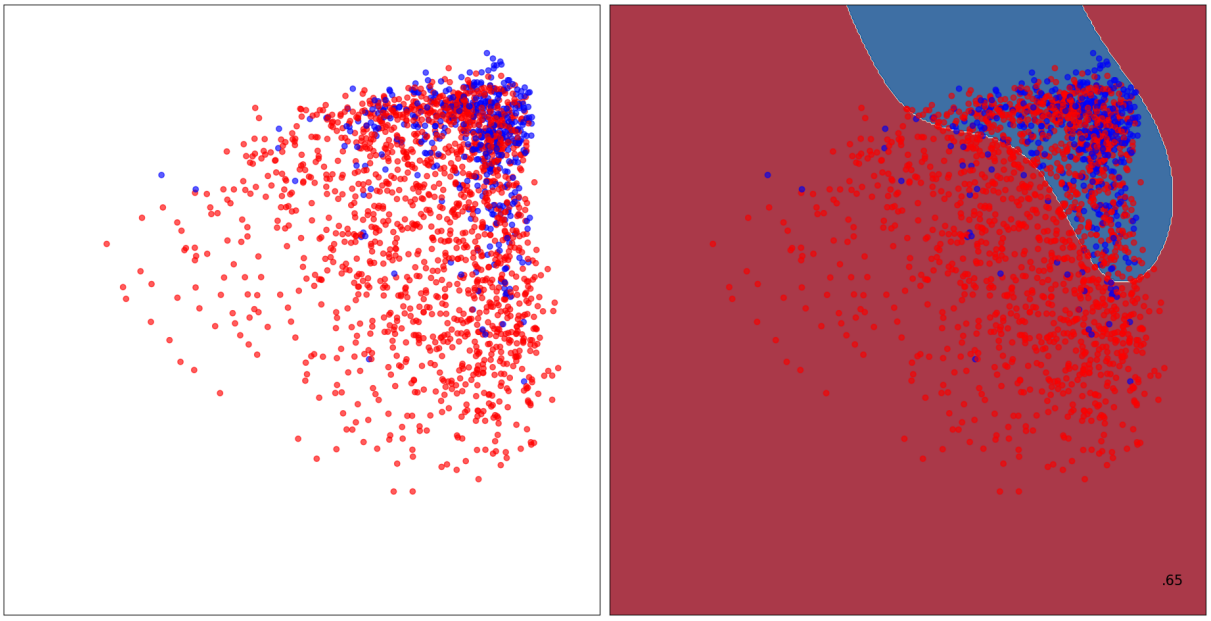

Figure 45 – SVM Liquids - /l/ and /r/

One can see here how the classification space is much more friendly with two classes, even in a two dimensional space. The liquids were classified at a high rate of 95% by the MLP and 96% by the SVM in the final models.

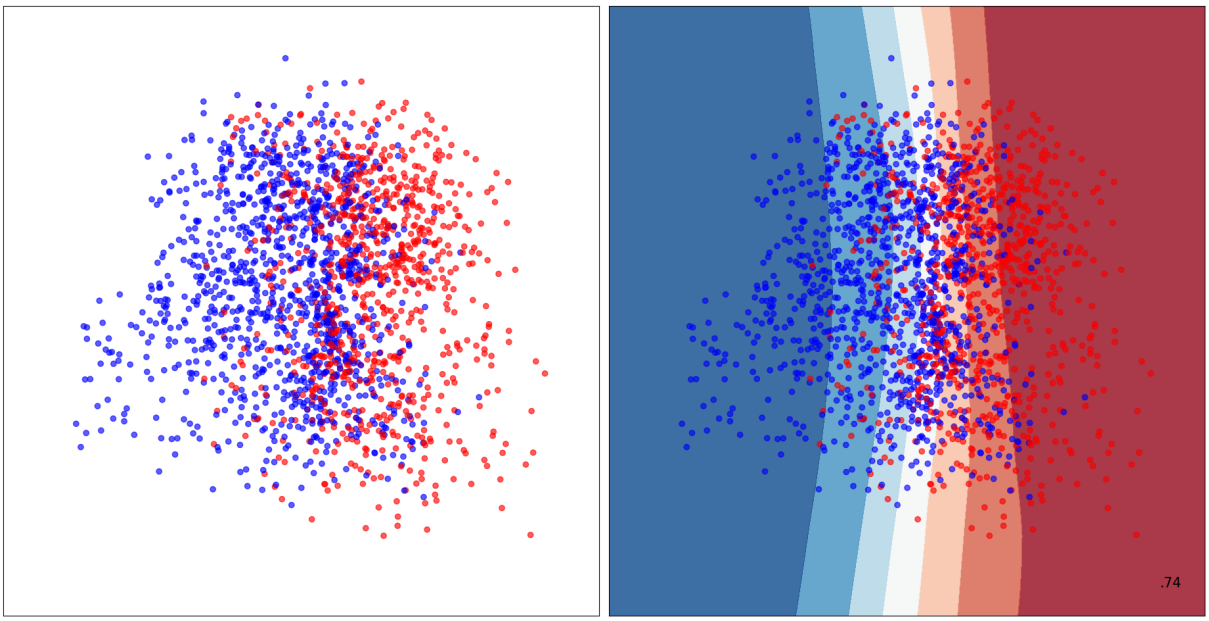

Figure 46 – MLP Glides - /y/ and /w/

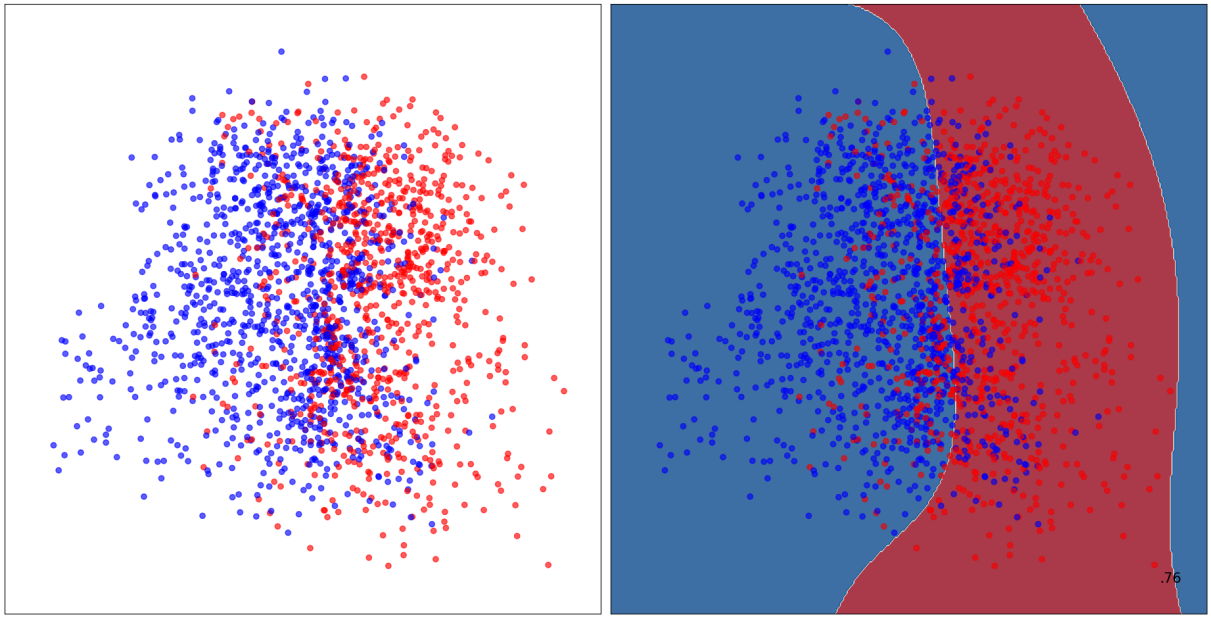

Figure 47 – SVM Glides - /y/ and /w/

The glides were even easier and in the end were classified 98% correctly by the SVM and 97% by the MLP. In a two dimensional space one can observe almost an even split down the middle.

Convergence Analysis

Here, we will show that the CNN converges on the RAMBLE data as it did on the TIMIT data. It should be noted that this formulation has been refined to be a more exact approximation. The formula used has been adapted from Mello, Ponti and Ferreira (2018) as:

$$\mathscr{N}(\mathscr{F}, 2n) \leq \sum_{i=0}^{\mathbb{E}(\mathrm{Cut}(A,B))} \binom{n}{i}$$

We can think of these cuts as edges on a hypothesis space, hypothetically a sphere which are cut by n hyperplanes. This formula calculates the angles of each vector which separates the space in order to optimize the point where the cut is made, given as:

$$\sum^{Cut(A,B)}$$

or, in other words, the largest number of edges (here estimated in terms of the expected value) cut by n within a space of m uniformly distributed features, a.k.a. the largest shattering coefficient for random vectors of m dimensions, the internal product, as proposed by Vapnik (2013)

In the case of the ramble dataset, we estimated:

$$\begin{aligned} N(F,2) \approx 2\sum_{i=0}^{598}\binom{n-1}{i}^{23} + \\ 2\sum_{i=0}^{598}\binom{n-1}{i}^{23} + \\ 2\sum_{i=0}^{910}\binom{n-1}{i}^{35} \end{aligned}$$

Remembering that the shattering coefficient is directly related to the number of dimensions and neurons in the case of the CNN, so we add all three layers together with their respective parameters.

$$\delta = 2N(F,2n)exp(-n \times 0.05^2/4) = 0.05)$$

So:

$$CNN(n) = 108,108,000,001,196$$

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{4/n(\log(2CNN(n)) - \log(0.05))}$$

which yields:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{4/8(\log(108,108,000,001,196^2) - \log(0.05))}$$

and:

$$\lim_{n\to\infty} \frac{\log\{CNN(n)\}}{n} \approx 0,$$

$$\lim_{n\to\infty} \frac{\log\{108,108,000,001,196\}}{n} \approx 0,$$

This means that 383,202 examples are needed for the CNN to converge where in practice we had 1,873,709.

5.3.2 Extrinsic Results

In the case of extrinsic evaluation, we want to see how well the model performs on a specific task. Here, the most appropriate task is the application for which this model was intended, namely, for use in a pronunciation tutor.

Epenthesis was chosen as the error to focus on since it is the error which most impedes communication for BP natives when speaking in English and because it is a particularly difficult construction for Brazilians (KLUGE *et al.*, 2007; MARTINS *et al.*, 2012; JOHN; CARDOSO, 2017). It is also interesting, because like most errors, it is rare that such epenthesis actually occurs, but when it does, the effect is quite obvious and undesirable. The unbalanced nature of this type of problem is also a good test for the model because it shows how robust it is. For the record, epenthesis was annotated to occur 50 times in the test set. For reference of the relative frequency, this error occured 83 in the test corpus used in Almeida (2016) which contained a total count of 2880 of all of the nine errors which were explained in the theoretical foundations. This means that epethesis has a relative frequency of about 2.88% of all errors in the corpus. Since much of the material in the ramble corpus is the same, these numbers should be very similar. While this number may seem small, it should be pointed out that vowel assimilations accounted for 47.43% of all errors and vowel nasalizations were present in 35.31%. This mean that these two errors already accounted for 82.74% of all errors in the corpus. The third most frequent of these error was consonant substitution which occurred in only 9.03% of the errors. This can be interpreted to show that the frequency of this error is not negligible and as mentioned earlier is well documented as an extremely difficult phenomenon for Brazilian L2 learners of English (KLUGE *et al.*, 2007; MARTINS *et al.*, 2012; JOHN; CARDOSO, 2017).

Real examples from the corpus can be seen in Table 20. The "@" symbol is used to denote the place at which the event occurs in relation to the word. While specific errors can be more rare than one might think, the particularly low value of 50 is probably due to more "careful" speech. Even though the participants were instructed to speak naturally, many of them spoke very slowly and consciously about their pronunciation since they were aware that the data would be used in a study about Brazilian English.

Table 20 – Examples of Epenthesis from the Corpus

Place	**Example**
Word Final	alice@
Word Initial	@almost
Interword	arriv@ed
Interword	call@ed
Interword	caus@ed
Interword	diagnos@ed
Interword	dri@ed
Interword	dropp@ed
Interword/Final	fac@ed@
Interword	forc@ed
Interword	help@ed
Interword	lauch@ed
Interword	liv@ed
Interword	look@ed
Word Initial	@made
Interword	mix@ed
Word Final	name@
Interword	nam@ed
Word Initial/Final	@nip@
Interword	observ@ed
Interword	pick@ed
Word Initial	@right
Interword	sal@esman
Word Initial	@scarlet
Word Initial	@screening
Interword	serv@ed
Word Initial	@small
Word Initial	@smaller
Word Initial	@sneezing
Interword	some@one
Word Initial	@spending
Word Initial	@state
Word Initial	@still
Interword	stopp@ed
Word Initial	@story
Word Initial	@straighten
Word Initial	@strain
Word Initial	@strike
Interword	talk@ed
Interword	us@ed
Interword	walk@ed
Word Final	wrote@

It is not surprising that the English past-tense ”-ed” morpheme is seen in a number of examples. The fact that it not only ends in a stop consonant, but is also written with a silent vowel letter is very difficult for BP natives and also quite difficult for a native English listener to understand when this error is produced.

The epenthesis identification results can be seen in Table 21. Here values are given for the identification of true positives as they occurred in the corpus. It should be noted that false positives are not applicable because since it is not an expected pronunciation, the absence of the error will not trigger an alarm. The difficulty is in recognizing that the

error does occur, especially in unexpected places.

Table 21 – Extrinsic Results for Epenthesis in Terms of Accuracy of True Positives

Place	**Accuracy**
All	88.0%
Word Initial	93.33%
Word Final	100%
Word Medial	83.33%

Table 21 shows how well the acoustic model is able to identify epenthesis in a task-based situation. These results are important because they demonstrate whether or not the model is actually able to perceive the insertion of the undesirable phoneme with a high confidence and segment it accordingly. This value represents a 69% relative gain over the result of 19.08% for true positives presented in (ALMEIDA, 2016) on a similar corpus, collected in the same way. It should also be noted that Almeida (2016) only counts ford initial and coda epenthesis and not word medial, which was more difficult for the acoustic model. It is possible that coda epenthesis is a mixture of word medial and word final epenthesis. In that study, the author reports a 12% accuracy in identifying word initial epenthesis and a 24% accuracy for coda epenthesis identification which proportionally seems correct when compared to the results in this dissertation.

In the case of a GMM-HMM model a score for possible phonemes in each frame is sent to a pronunciation model where the heavy lifting is done. This is an efficient strategy in order to identify the correct words in a sentence but it is detrimental to phoneme recognition because unless the non-canonical or “erroneous” word is actually in the dictionary, the identification of this type of error is impossible.

In the case of a neural model, the acoustic model is bounded to the language and pronunciation model in an end to end approach. This is much more efficient considering the huge search beam space available to the model. The problem with this approach, in the case of phoneme recognition, is that we no longer get phoneme recognition results but rather the inputs and outputs for finite state transducers which are highly efficient in bridging the gap between the pronunciation and language model but obtaining the exact posterior value for phonemes is not quite straight forward. Again, this approach is highly dependent on the resources at hand. Also, in the case where the model is used to make frame or boundary-wise predictions, the results are probabilistic and not robust enough to generalize well for disfluencies since they occur rarely.

In the case of the HMM boundaries predicted by the CNN-HTSVM in this thesis, it is possible to claim that predictions can be well inferred based on solid learning guarantees and that an accurate feature space can be separated in a way which would be useful for the detection of pronunciation errors.

CHAPTER

6

CONCLUSIONS

First, it is time to sum up the research questions made in this thesis:

As mentioned earlier, two research questions must be answered:

1. Does the acoustic model for accented speech achieve results which are better than the SotA and bring us closer to the results produced by manual alignment?

   The results presented here are better than or near the results for non-native speech with very large corpora. With less than a tenth of the data robust results were obtained on a much more difficult corpus. Garber, Singer and Ward (2017) used 800 hours in domain speech for aircrafts. The model presented in this dissertation is a general model and contains a variety of noise in the data. Also, even with more overlapping classes and increased, only a 10 point reduction was seen in the method as compared to the experiments on the TIMIT dataset. This further evidences that the classes were able to be separated even with little data. In the case of native speech, the results are not better than the SotA and are not a new baseline. Still, the results are not far from the SotA for native speech but still there remains a gap for non-native speech. Most importantly, the work presented in this dissertation does not require external resources as the SotA approaches do. As far as reproducible methods which make no use of external datasets or weights, these are likely the best framewise phoneme recognition results known in the literature to date for non-native speech. Most importantly, non-native and noisy speech recognition can achieve good results without large amounts of data.

2. Is the model robust?

   The model is intrinsically robust because it provides solid learning guarantees af-

forded by the statistical learning theory. It is also extrinsically robust because it is able to identify the occurrence of epenthesis by Brazilian speaker with high accuracy. An intelligent speech tutor for Brazilian learners of English, capable of treating errors like epenthesis, would be possible using this method.

6.1 Main Contributions

The main contribution of this thesis was the proposal of a transparent method with solid statistical learning guarantees. The method has proven to be intrinsically robust with strong F-measure scores and proven convergence as well as extrinsically robust in a task-based scenario. A third contribution is a method which is robust in adverse conditions like noisy data and the confusion generated by a foreign accent. Last but not least, this dissertation shows that good results for speech recognition can be obtained using deep feature extraction and without big data. All of these contributions are relevant to SotA in Speech recognition as the technology matures and enters everyday life.

6.2 Limitations

The scope of this thesis was focused on the acoustic model. A full ASR engine would have been interesting, but unfortunately it could not fit into the scope of the thesis. With a robust acoustic model for accented speech it will be possible to develop a complete intelligent pronunciation tutor for all nine mispronunciation errors presented in the Introduction Section.

The greatest restriction at this point is the accuracy of language independent non-native speech in ASR systems with unrestricted speech. This would be the ideal situation. Even with restricted speech this is not a trivial task. Another restriction due to time is the types of non-native speech markers which could be treated. It would be important to be able to provide information on prosody, like the pitch contour for example, but this is just outside of the scope of an acoustic model.

6.3 Future Work

In the future it would be interesting to further explore the capabilities of the CNN as a feature extractor. Perhaps audio-based features could also offer a multi-modal approach. Some work on the HTSVM would be useful, especially in cases where it ends up on the wrong side of the tree. Some kind of back-tracking algorithm could help in that case. One of the greatest difficulties was in discriminating vowels. This could probably be handled in two ways: 1) a comparison of errors to those actually occurring in the corpus

data to check the annotations; and 2.) adding more features to better separate very close classes.

6.4 Conclusions

It is important to note that the results produced in this thesis are reproducible and no pretraining or fine-tuning is needed to achieve comparable results to the SotA. The closest reproduction paper by Niedek, Heskes and Leeuwen (2016) still uses TensorFlow's default weights. Graves, Jaitly and Mohamed (2013) and Song and Cai (2015) are still the industry benchmarks but without proprietary data, it is much more difficult to understand how these networks actually work. Large networks like this simply cannot generalize well for small datasets and currently rely on heavy regularization and fine-tuning. This is important because it means that we can generalize well only with the small amount of data available. Fine-tuning is also detrimental to generalization. While this practice is useful for quickly optimizing the training iterations for a specific data set, it is not useful when transferred to another dataset or when new noise is present.

The problem here is not whether or not a benchmark can be beaten but what to do when conditions change. In the big data scenario, methods like MCT (KIM *et al.*, 2016), a typical way to compensate by multiplying the data for all necessary variations, seems impossible for any low resource scenario and not feasible even when data is readily available.

Another important outcome of this thesis were the window experiments where, like other researchers, we found that 11 frames seems to be the most cost-efficient number, since larger configurations demand a much larger number of feature maps and offer little improvement. It would be interesting to test this hypothesis on non-European languages to see if this still holds true.

The take home point of this dissertation is that a quite robust acoustic model can be built even with a simple architecture and small dataset when carefully constructed. Robust means that the results are similar to related recent studies with a 37.04% FER and the convergence analysis provides sufficient evidence that the model will infer unseen data well, guaranteeing that the results were not obtained by chance.

Other outcomes include the resources produced. It is the hoped that the speech corpus produced in this thesis can be used for a wide variety of speech experiments and hopefully will continue to grow in size and quality. More annotation with other linguistic futures would be useful, especially in the case of prosody. The focus here was on phoneme classification but it is well known that prosody contributes a great deal to the accentedness of speech.

Beyond the corpus itself, several tools were used to produce it like the phonetic balancer scripts as well as the PRAAT annotation plugin. These tools should be excellent contributions to the community and hopefully will be adopted by the industry and academia and better still, I would welcome any issues or extensions to those codes (all the resources and code are publicly available in github).

6.5 Technical Production

A series of programs and source code was written for this dissertation. Here is a list of each item and its location, free for public use under GNU public license v.3.0:

The Ramble Speech Corpus: https://github.com/CShulby/Ramble-Corpus

The Ramble Acoustic Models:
https://github.com/CShulby/Ramble-Acoustic-Models

The htklabel plugin from the CPrAN repository at:
http://cpran.net/plugins/htklabel/

or the bleeding edge version:
https://github.com/CShulby/plugin_htklabel/

Additionally, a version for speech synthesis (HTS labels) was also created:
https://github.com/CShulby/plugin_htslabel/

The Greedy Corpus Balancer Script:
https://github.com/CShulby/Balancer-Scripts

6.6 Scientific Production

This section lists all of the publications in academic journals and conferences. Although not all of them are directly related to this dissertation they are related to NLP in general.

Shulby, C., Pombal, L., Ziolle, G., Jordão, V., Mathos, B., Postal, A., & Prochnow, T. (Accepted). Proactive Security: Embedded AI Solution for Violent and Abusive Speech Recognition. In Proceedings of the 12th Brazilian Symposium in Information and Human

Language Technology.

Treviso, M. V., dos Santos, L. B., Shulby, C., Hübner, L. C., Mansur, L. L., & Aluísio, S. M. (2018). Detecting mild cognitive impairment in narratives in Brazilian Portuguese: first steps towards a fully automated system. Letras de Hoje, 53(1), 48-58.

Shulby, C. D., Ferreira, M. D., de Mello, R. F., & Aluisio, S. M. (2017). Acoustic Modeling Using a Shallow CNN-HTSVM Architecture. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (pp. 85-90).

Treviso, M., Shulby, C., & Aluísio, S. (2017). Evaluating Word Embeddings for Sentence Boundary Detection in Speech Transcripts. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (pp. 151-160).

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Silva, J., & Aluísio, S. (2017). Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (pp. 122-131).

Treviso, M., Shulby, C., & Aluísio, S. (2017). Sentence Segmentation in Narrative Transcripts from Neuropsychological Tests using Recurrent Convolutional Neural Networks. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (Vol. 1, pp. 315-325).

Marquiafável, V., Shulby, C., Veiga, A., Proença, J., Candeias, S., & Perdigão, F. (2014). Rule-Based Algorithms for Automatic Pronunciation of Portuguese Verbal Inflections. In Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014, Proceedings (Vol. 8775, p. 36). Springer.

Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., & Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. In Telecommunications Symposium (ITS), 2014 International (pp. 1-5). IEEE.

Shulby, C., Mendonça, G., & Marquiafável, V. (2013). Automatic disambiguation of homographic heterophone pairs containing open and closed mid vowels. In Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology.

BIBLIOGRAPHY

ABDEL-HAMID, O.; MOHAMED, A.-r.; JIANG, H.; PENN, G. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In: IEEE. **2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)**. [S.l.], 2012. p. 4277–4280. Citations on pages 70, 72, 77, 101, and 103.

ABDEL-HAMID, O.; MOHAMED, A.-R.; JIANG, H.; DENG, L.; PENN, G.; YU, D. Convolutional neural networks for speech recognition. **IEEE/ACM Transactions on audio, speech, and language processing**, IEEE, v. 22, n. 10, p. 1533–1545, 2014. Citations on pages 17, 33, 50, 59, 70, 72, 77, 88, 101, 103, and 118.

ALISSON, E. Usp de são carlos inaugura supercomputador. **Agência FAPESP**, 2015. Accessed: 2018-03-01. Citation on page 73.

ALMEIDA, G. A. d. M. **Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2016. Citations on pages 17, 30, 40, 41, 42, 90, 137, and 139.

AMAMI, R.; ELLOUZE, N. Study of phonemes confusions in hierarchical automatic phoneme recognition system. **arXiv preprint arXiv:1508.01718**, 2015. Citations on pages 79, 80, 112, and 115.

BAUER, D. d. A.; ALVES, U. K. O ensino comunicativo de pronúncia nas aulas de inglês (l2) para aprendizes brasileiros: análise de um livro didático. **Revista Linguagem & Ensino**, v. 14, n. 2, p. 287–314, 2012. Citations on pages 26 and 27.

BERGER, C. R.; CALABRESE, R. J. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. **Human communication research**, Wiley Online Library, v. 1, n. 2, p. 99–112, 1975. Citation on page 28.

BLAIR, L. **Straight Talking: Learn to overcome insomnia, anxiety, negative thinking and other modern day stresses.** [S.l.]: CNIB, 2013. Citation on page 28.

BOERSMA, P. Praat: doing phonetics by computer. **http://www. praat. org/**, 2006. Citations on pages 60 and 82.

CANAVAN, A.; GRAFF, D.; ZIPPERLEN, G. Callhome american english speech. **Linguistic Data Consortium**, 1997. Citation on page 72.

CARLETTA, J. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. **Language Resources and Evaluation**, Springer, v. 41, n. 2, p. 181–190, 2007. Citation on page 73.

CELCE-MURCIA, M.; BRINTON, D. M.; GOODWIN, J. M. **Teaching pronunciation hardback with audio CDs (2): A course book and reference guide**. [S.l.]: Cambridge University Press, 2010. Citation on page 38.

CERNAK, M.; IMSENG, D.; BOURLARD, H. Robust triphone mapping for acoustic modeling. In: **Proceedings of Interspeech**. [S.l.: s.n.], 2012. Citation on page 74.

CHANG, Y.-W.; HSIEH, C.-J.; CHANG, K.-W.; RINGGAARD, M.; LIN, C.-J. Training and testing low-degree polynomial data mappings via linear svm. **Journal of Machine Learning Research**, v. 11, n. Apr, p. 1471–1490, 2010. Citation on page 103.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002. Citation on page 105.

CHEN, S. F.; GOODMAN, J. An empirical study of smoothing techniques for language modeling. **Computer Speech & Language**, Elsevier, v. 13, n. 4, p. 359–394, 1999. Citation on page 53.

CHIU, C.-C.; SAINATH, T. N.; WU, Y.; PRABHAVALKAR, R.; NGUYEN, P.; CHEN, Z.; KANNAN, A.; WEISS, R. J.; RAO, K.; GONINA, K. *et al.* State-of-the-art speech recognition with sequence-to-sequence models. **arXiv preprint arXiv:1712.01769**, 2017. Citation on page 72.

CHOLLET, F. *et al.* **Keras**. 2015. <https://keras.io>. Citation on page 102.

CHOMSKY, N. **Aspects of the Theory of Syntax**. [S.l.]: MIT press, 2014. Citation on page 53.

CHUN, D. M. Computer-assisted pronunciation teaching. **The Encyclopedia of Applied Linguistics**, Wiley Online Library, 2012. Citations on pages 25 and 27.

COATES, A.; HUVAL, B.; WANG, T.; WU, D.; CATANZARO, B.; ANDREW, N. Deep learning with cots hpc systems. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2013. p. 1337–1345. Citation on page 73.

CRISTÓFARO, T. Pronúncia do inglês: para falantes do português brasileiro. **Rio de Janeiro: Contexto**, 2015. Citations on pages 26, 27, 30, 37, and 40.

CRUZ, M. d. L. O. B. **Etapas de interlengua oral en estudiantes brasileños de español**. [S.l.]: Asele, 2004. Citations on pages 31 and 36.

DAHL, G. E.; YU, D.; DENG, L.; ACERO, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. **IEEE Transactions on audio, speech, and language processing**, IEEE, v. 20, n. 1, p. 30–42, 2012. Citations on pages 17 and 52.

DAM, P. T. Mother-tongue interference in spanish-speaking english language learners' interlanguage. **English Language Learners in 21st Century Classrooms: Challenges and Expectations**, Texas Woman s University, 2012. Citations on pages 31 and 36.

DEKEL, O.; KESHET, J.; SINGER, Y. An online algorithm for hierarchical phoneme classification. In: SPRINGER. **International Workshop on Machine Learning for Multimodal Interaction**. [S.l.], 2004. p. 146–158. Citation on page 79.

DEMENKO, G.; WAGNER, A.; CYLWIK, N. The use of speech technology in foreign language pronunciation training. **Archives of Acoustics**, v. 35, n. 3, p. 309–329, 2010. Citation on page 26.

DERWING, T. What do esl students say about their accents? **Canadian Modern Language Review**, University of Toronto Press, v. 59, n. 4, p. 547–567, 2003. Citation on page 28.

DRIAUNYS, K.; RUDŽIONIS, V.; ŽVINYS, P. Implementation of hierarchical phoneme classification approach on ltdigits corpora. **Information Technology and Control**, v. 38, n. 4, 2015. Citations on pages 79 and 80.

EGAN, K. B. Speaking: A critical skill and a challenge. **Calico Journal**, JSTOR, p. 277–293, 1999. Citation on page 25.

EISENSTEIN, M. Native reactions to non-native speech: A review of empirical research. **Studies in Second Language Acquisition**, Cambridge University Press, v. 5, n. 2, p. 160–176, 1983. Citation on page 37.

ELLIS, R. Variability and the natural order hypothesis. **Beyond the monitor model**, p. 139–158, 1994. Citations on pages 31 and 36.

Fernandes de Mello, R.; Antonelli Ponti, M.; Grossi Ferreira, C. H. Computing the Shattering Coefficient of Supervised Learning Algorithms. **ArXiv e-prints**, May 2018. Citations on pages 114 and 115.

Fernandes de Mello, R.; Dais Ferreira, M.; Antonelli Ponti, M. Providing theoretical learning guarantees to Deep Learning Networks. **ArXiv e-prints**, 2017. Citations on pages 66 and 113.

FERREIRA, M. D.; CORREA, D. C.; NONATO, L. G.; MELLO, R. F. de. Designing architectures of convolutional neural networks to solve practical problems. **2017 Elesvier pre-print**, 2017. Citation on page 117.

FERREIRA, M. D.; CORRêA, D. C.; NONATO, L. G.; MELLO, R. F. de. Designing architectures of convolutional neural networks to solve practical problems. **Expert Systems with Applications**, v. 94, n. Supplement C, p. 205 – 217, 2018. ISSN 0957-4174. Citations on pages 101, 110, and 114.

FIRST, E. E. **EF EPI: EF English Proficiency Index**. 2017. Available: <https://www.ef.com.br/epi/>. Citations on pages 26 and 28.

FISCUS, J.; GAROFOLO, J.; PRZYBOCKI, M.; FISHER, W.; PALLETT, D. English broadcast news speech (hub4). **Linguistic Data Consortium, Philadelphia**, 1997. Citation on page 72.

FLEGE, J. E.; MUNRO, M. J.; MACKAY, I. R. Factors affecting strength of perceived foreign accent in a second language. **The Journal of the Acoustical Society of America**, ASA, v. 97, n. 5, p. 3125–3134, 1995. Citations on pages 26 and 28.

FORSBERG, M. Why is speech recognition difficult. **Chalmers University of Technology**, 2003. Citation on page 74.

FROMKIN, V.; KRASHEN, S.; CURTISS, S.; RIGLER, D.; RIGLER, M. The development of language in genie: a case of language acquisition beyond the "critical period". **Brain and language**, Elsevier, v. 1, n. 1, p. 81–107, 1974. Citation on page 36.

FUERTES, J. N.; POTERE, J. C.; RAMIREZ, K. Y. Effects of speech accents on interpersonal evaluations: implications for counseling practice and research. **Cultural Diversity and Ethnic Minority Psychology**, Educational Publishing Foundation, v. 8, n. 4, p. 346, 2002. Citation on page 37.

FUND, I. M. World economic outlook database. **World Econ Finance Survey**, 2017. Citation on page 25.

_____. World economic outlook. **Accessed from: http://www.imf.org/en/Publications/WEO/Issues/2018/01/11/world-economic-outlook-update-january-2018**, 2018. Citation on page 26.

GALES, M.; YOUNG, S. *et al.* The application of hidden markov models in speech recognition. **Foundations and Trends® in Signal Processing**, Now Publishers, Inc., v. 1, n. 3, p. 195–304, 2008. Citations on pages 17, 29, 54, and 71.

GARBER, M.; SINGER, M.; WARD, C. Accent adaptation for the air traffic control domain. In: **Proceedings of ACL 2017, Student Research Workshop**. [S.l.: s.n.], 2017. p. 95–99. Citations on pages 77, 120, 121, and 141.

GAROFOLO, J.; LAMEL, L.; FISHER, W.; FISCUS, J.; PALLETT, D.; DAHLGREN, N. The darpa timit acoustic-phonetic continuous speech corpus, ntis speech disc. **NTIS order number PB91-100354**, 1990. Citations on pages 96 and 97.

GASS, S. M. **Second language acquisition: An introductory course**. [S.l.]: Routledge, 2013. Citations on pages 36 and 40.

GENÇ, B.; BADA, E. English as a world language in academic writing. **Reading Matrix: An International Online Journal**, v. 10, n. 2, 2010. Citation on page 27.

GENT, E. Ai: Fears of'playing god'. **Engineering & Technology**, IET, v. 10, n. 2, p. 76–79, 2015. Citations on pages 71 and 73.

GERS, F. A.; SCHRAUDOLPH, N. N.; SCHMIDHUBER, J. Learning precise timing with lstm recurrent networks. **Journal of machine learning research**, v. 3, n. Aug, p. 115–143, 2002. Citations on pages 17 and 57.

GILAKJANI, A. P. The significance of pronunciation in english language teaching. **English language teaching**, Canadian Center of Science and Education (CCSE), v. 5, n. 4, p. 96, 2012. Citation on page 25.

GODFREY, J. J.; HOLLIMAN, E. Switchboard-1. **Linguistic Data Consortium**, 1993. Citation on page 72.

GOLDBERG, Y.; ELHADAD, M. splitsvm: fast, space-efficient, non-heuristic, polynomial kernel computation for nlp applications. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers**. [S.l.], 2008. p. 237–240. Citation on page 103.

GOLDMAN, J.-P. Easyalign: an automatic phonetic alignment tool under praat. 2011. Citation on page 82.

GOODFELLOW YOSHUA BENGIO, A. C. I. Deep learning. Book in preparation for MIT Press. 2016. Available: <http://goodfeli.github.io/dlbook/>. Citation on page 101.

GRAHAM, W. **Phonological Processes**. 2014. <https://www.sltinfo.com/phonological-processes/>. Accessed: 2018-06-01. Citation on page 40.

GRAVES, A.; JAITLY, N. Towards end-to-end speech recognition with recurrent neural networks. In: **ICML**. [S.l.: s.n.], 2014. v. 14, p. 1764–1772. Citations on pages 70 and 101.

GRAVES, A.; JAITLY, N.; MOHAMED, A.-r. Hybrid speech recognition with deep bidirectional lstm. In: **Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on**. Piscataway, NJ: IEEE, 2013. p. 273–278. Citations on pages 84, 85, 88, and 143.

GRAVES, A.; LIWICKI, M.; FERNÁNDEZ, S.; BERTOLAMI, R.; BUNKE, H.; SCHMIDHUBER, J. A novel connectionist system for unconstrained handwriting recognition. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 31, n. 5, p. 855–868, 2009. Citation on page 56.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. **Acoustics, speech and signal processing (icassp), 2013 ieee international conference on**. [S.l.], 2013. p. 6645–6649. Citations on pages 78, 84, and 85.

GRAVES, A.; SCHMIDHUBER, J. Framewise phoneme classification with bidirectional lstm and other neural network architectures. **Neural Networks**, Elsevier, v. 18, n. 5-6, p. 602–610, 2005. Citation on page 119.

HAMMARBERG, B. Conditions on transfer in phonology. In: JAMES, J. L. A. R. (Ed.). **Second-language Speech: Structure and Process**. first. [S.l.]: Mouton de Gruyter, 1997. p. 161–180. Citation on page 37.

HANNUN, A. Y.; MAAS, A. L.; JURAFSKY, D.; NG, A. Y. First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns. **arXiv preprint arXiv:1408.2873**, 2014. Citation on page 70.

HAU, D.; CHEN, K. Exploring hierarchical speech representations with a deep convolutional neural network. **UKCI 2011 Accepted Papers**, p. 37, 2011. Citation on page 70.

HERNANDEZ, D. The man behind the google brain: Andrew ng and the quest for the new ai. **Wired),[online] May**, v. 7, 2013. Citations on pages 71 and 73.

HIFNY, Y.; RENALS, S. Speech recognition using augmented conditional random fields. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 17, n. 2, p. 354–365, 2009. Citation on page 83.

HINCKS, R. Speech technologies for pronunciation feedback and evaluation. **ReCALL**, Cambridge University Press, v. 15, n. 1, p. 3–20, 2003. Citations on pages 25 and 27.

HINTON, G.; DENG, L.; YU, D.; DAHL, G. E.; MOHAMED, A.-r.; JAITLY, N.; SENIOR, A.; VANHOUCKE, V.; NGUYEN, P.; SAINATH, T. N. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal Processing Magazine**, IEEE, v. 29, n. 6, p. 82–97, 2012. Citations on pages 29, 70, 71, 72, and 117.

HINTON, G. E. Connectionist learning procedures. **Artif. Intell.**, Elsevier Science Publishers Ltd., Essex, UK, v. 40, n. 1-3, p. 185–234, Sep. 1989. ISSN 0004-3702. Available: <http://dx.doi.org/10.1016/0004-3702(89)90049-0>. Citation on page 110.

HOCHREITER, S. Untersuchungen zu dynamischen neuronalen netzen [in german] diploma thesis. **TU Münich**, 1991. Citation on page 57.

HOFBAUER, K.; PETRIK, S.; HERING, H. The atcosim corpus of non-prompted clean air traffic control speech. In: **LREC**. [S.l.: s.n.], 2008. Citation on page 77.

HOFFMANN, S.; TIK, E. Automatic phone segmentation. **Corpora**, v. 3, p. 2–1, 2009. Citation on page 115.

HUANG, Z.; ZWEIG, G.; DUMOULIN, B. Cache based recurrent neural network language model inference for first pass speech recognition. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on**. [S.l.], 2014. p. 6354–6358. Citation on page 72.

HUBEL, D. H.; WIESEL, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. **The Journal of physiology**, Wiley Online Library, v. 160, n. 1, p. 106–154, 1962. Citation on page 58.

HYLTENSTAM, K.; ABRAHAMSSON, N. Who can become native-like in a second language? all, some, or none? **Studia linguistica**, Wiley Online Library, v. 54, n. 2, p. 150–166, 2000. Citation on page 37.

IVERSON, P.; KUHL, P. K.; AKAHANE-YAMADA, R.; DIESCH, E.; TOHKURA, Y.; KETTERMANN, A.; SIEBERT, C. A perceptual interference account of acquisition difficulties for non-native phonemes. **Cognition**, Elsevier, v. 87, n. 1, p. B47–B57, 2003. Citations on pages 26 and 28.

JAITLY, N.; NGUYEN, P.; SENIOR, A.; VANHOUCKE, V. Application of pretrained deep neural networks to large vocabulary speech recognition. In: **Thirteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2012. Citation on page 72.

JIAMPOJAMARN, S.; KONDRAK, G.; SHERIF, T. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In: **Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference**. Rochester, New York: Association for Computational Linguistics, 2007. p. 372–379. Available: <http://www.aclweb.org/anthology/N/N07/N07-1047>. Citation on page 96.

JOHN, P.; CARDOSO, W. On syllable structure and phonological variation: The case of i-epenthesis by brazilian portuguese learners of english. **Ilha do Desterro**, SciELO Brasil, v. 70, n. 3, p. 169–184, 2017. Citations on pages 32, 107, and 137.

JUAN, S. S.; BESACIER, L.; LECOUTEUX, B.; TAN, T.-P. Merging of native and non-native speech for low-resource accented asr. In: SPRINGER. **International Conference on Statistical Language and Speech Processing**. [S.l.], 2015. p. 255–266. Citation on page 75.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. [S.l.]: Pearson London:, 2014. Citation on page 52.

KARPAGAVALLI, S.; CHANDRA, E. A hierarchical approach in tamil phoneme classification using support vector machine. **Indian Journal of Science and Technology**, v. 8, n. 35, 2015. Citations on pages 79, 80, 111, and 116.

KIM, T. Y.; HAN, C. W.; KIM, S.; AHN, D.; JEONG, S.; LEE, J. W. Korean lvcsr system development for personal assistant service. In: IEEE. **Consumer Electronics (ICCE), 2016 IEEE International Conference on**. [S.l.], 2016. p. 93–96. Citations on pages 105 and 143.

KLUGE, D. C.; RAUBER, A. S.; REIS, M. S.; BION, R. A. H. The relationship between the perception and production of english nasal codas by brazilian learners of english. In: **Eighth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2007. Citations on pages 32 and 137.

KRASHEN, S. D. **The input hypothesis: Issues and implications**. [S.l.]: Addison-Wesley Longman Ltd, 1985. Citation on page 37.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2012. p. 1097–1105. Citation on page 58.

LABOV, W. The social motivation of a sound change. **Word**, Taylor & Francis, v. 19, n. 3, p. 273–309, 1963. Citation on page 28.

LABOV, W.; ASH, S.; BOBERG, C. **The atlas of North American English: Phonetics, phonology and sound change**. [S.l.]: Walter de Gruyter, 2005. Citation on page 31.

LADEFOGED, P.; DISNER, S. F. **Vowels and consonants**. [S.l.]: John Wiley & Sons, 2012. Citations on pages 33, 74, 103, and 104.

LADEFOGED, P.; JOHNSON, K. **A course in phonetics**. [S.l.]: Nelson Education, 2014. Citations on pages 28, 61, and 62.

LANG, Y.; WANG, L.; SHEN, L.; WANG, Y. An integrated approach to the teaching and learning of zh. **Electronic Journal of Foreign Language Teaching**, v. 9, n. 2, 2012. Citation on page 28.

LECUN, Y.; BENGIO, Y. Convolutional networks for images, speech, and time series. **The handbook of brain theory and neural networks**, v. 3361, n. 10, p. 1995, 1995. Citations on pages 101 and 103.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Citation on page 101.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, IEEE, v. 86, n. 11, p. 2278–2324, 1998. Citations on pages 50 and 101.

LEE, H.; PHAM, P.; LARGMAN, Y.; NG, A. Y. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2009. p. 1096–1104. Citation on page 70.

LEE, K.-F.; HON, H.-W. Speaker-independent phone recognition using hidden markov models. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, IEEE, v. 37, n. 11, p. 1641–1648, 1989. Citation on page 97.

LEI, X.; SENIOR, A. W.; GRUENSTEIN, A.; SORENSEN, J. Accurate and compact large vocabulary speech recognition on mobile devices. In: **Interspeech**. [S.l.: s.n.], 2013. v. 1. Citations on pages 29, 71, and 72.

LENNEBERG, E. H. The biological foundations of language. **Hospital Practice**, Taylor & Francis, v. 2, n. 12, p. 59–67, 1967. Citation on page 36.

LEVENSHTEIN, V. I. **Binary codes capable of correcting deletions, insertions, and reversals**. [S.l.], 1966. v. 10, n. 8, 707–710 p. Citations on pages 83 and 111.

LIAO, H.; MCDERMOTT, E.; SENIOR, A. Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription. In: IEEE. **Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on**. [S.l.], 2013. p. 368–373. Citation on page 73.

LOMBART, J.; MIGUEL, A.; LLEIDA, E. Articulatory feature extraction from voice and their impact on hybrid acoustic models. In: **Advances in Speech and Language Technologies for Iberian Languages**. [S.l.]: Springer, 2014. p. 138–147. Citations on pages 85, 111, 116, and 119.

LOPES, C.; PERDIGÃO, F. *et al.* Phonetic recognition improvements through input feature set combination and acoustic context window widening. In: CITESEER. **7th Conference on Telecommunications, Conftele**. [S.l.], 2009. p. 449–452. Citations on pages 86, 110, 118, and 119.

LUXBURG, U. V.; SCHÖLKOPF, B. Statistical learning theory: Models, concepts, and results. **arXiv preprint arXiv:0810.4752**, 2008. Citations on pages 17, 33, 64, 66, 103, 111, 113, 114, and 115.

LYSTER, R. Roles for corrective feedback in second language instruction. **The Encyclopedia of Applied Linguistics**, Wiley Online Library, 2013. Citations on pages 25 and 38.

_____. Using form-focused tasks to integrate language across the immersion curriculum. **System**, Elsevier, v. 54, p. 4–13, 2015. Citation on page 38.

MACLEAN, K. **Tutorial: Create Acoustic Model - Manually**. 2018. Accessed: 2018-03-01. Available: <http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial/triphones/step-10>. Citations on pages 101 and 110.

MAKHOUL, J.; SCHWARTZ, R. State of the art in continuous speech recognition. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 92, n. 22, p. 9956–9963, 1995. Citation on page 71.

MARTIN, J. H.; JURAFSKY, D. Speech and language processing: An introduction to natural language processing. **Computational Linguistics and Speech Recognition. Prentice Hall**, v. 2, 2000. Citation on page 71.

MARTINS, D. H. *et al.* The production of english final by brazilians. Florianópolis, 2012. Citations on pages 32 and 137.

MASSARO, D. W.; COHEN, M. M. Phonological context in speech perception. **Perception & Psychophysics**, v. 34, n. 4, p. 338–348, Jul 1983. ISSN 1532-5962. Available: <https://doi.org/10.3758/BF03203046>. Citation on page 39.

MCCROCKLIN, S. M. **The potential of Automatic Speech Recognition for fostering pronunciation learners' autonomy**. Phd Thesis (PhD Thesis) — Iowa State University, 2014. Citation on page 28.

MELLO, R. F. de; PONTI, M. A.; FERREIRA, C. H. G. Computing the shattering coefficient of supervised learning algorithms. **arXiv preprint arXiv:1805.02627**, 2018. Citation on page 135.

MENDONÇA, G.; CANDEIAS, S.; PERDIGÃO, F.; SHULBY, C.; TONIAZZO, R.; KLAUTAU, A.; ALUÍSIO, S. A method for the extraction of phonetically-rich triphone sentences. In: IEEE. **Telecommunications Symposium (ITS), 2014 International**. [S.l.], 2014. p. 1–5. Citations on pages 90 and 93.

MEZA, I. **Identificacion de hablante**. 2018. <http://turing.iimas.unam.mx/~ivanvladimir/slides/fonologia_forense/identification_cont.html>. Accessed: 2018-06-01. Citations on pages 17 and 49.

MIKOLOV, T.; KARAFIÁT, M.; BURGET, L.; ČERNOCKỲ, J.; KHUDANPUR, S. Recurrent neural network based language model. In: **Eleventh Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2010. Citations on pages 53 and 56.

MITCHELL, T. M. *et al.* **Machine learning. WCB**. [S.l.]: McGraw-Hill Boston, MA:, 1994. Citation on page 58.

MOHAMED, A.-r.; DAHL, G. E.; HINTON, G. Acoustic modeling using deep belief networks. **IEEE Transactions on Audio, Speech, and Language Processing**, IEEE, v. 20, n. 1, p. 14–22, 2012. Citations on pages 33 and 70.

MUNRO, M. J.; DERWING, T. M.; SATO, K. *et al.* Salient accents, covert attitudes: Consciousness-raising for pre-service second language teachers. AMEP Research Centre, 2006. Citations on pages 26, 28, and 37.

NIEDEK, T. van; HESKES, T.; LEEUWEN, D. van. Phonetic classification in tensorflow. **Bachelor's Thesis, Radboud University**, 2016. Citations on pages 84 and 143.

ODDEN, D. **Introducing phonology**. [S.l.]: Cambridge university press, 2005. Citation on page 74.

OECD. **PISA 2015 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science**. 2015. Available: <http://www.oecd.org/pisa/PISA-2015-Brazil-PRT.pdf>. Citation on page 26.

OORD, A. V. D.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K. Wavenet: A generative model for raw audio. **arXiv preprint arXiv:1609.03499**, 2016. Citation on page 50.

PASCANU, R.; MIKOLOV, T.; BENGIO, Y. Understanding the exploding gradient problem. **CoRR, abs/1211.5063**, 2012. Citation on page 74.

PEARSON. **Heightened Urgency for Business English in an Increasingly Global Workforce: A look at the 2013 Business English Index & Globalization of English Report**. 2014. Available: <http://static.globalenglish.com/files/reports/Business_English_Index_2013.pdf>. Citation on page 28.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citation on page 110.

PITT, M. A.; DILLEY, L.; JOHNSON, K.; KIESLING, S.; RAYMOND, W.; HUME, E.; FOSLER-LUSSIER, E. Buckeye corpus of conversational speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. **Ohio State University (Distributor)**, 2007. Citation on page 47.

POVEY, D.; GHOSHAL, A.; BOULIANNE, G.; BURGET, L.; GLEMBEK, O.; GOEL, N.; HANNEMANN, M.; MOTLICEK, P.; QIAN, Y.; SCHWARZ, P. *et al.* The kaldi speech recognition toolkit. In: IEEE SIGNAL PROCESSING SOCIETY. **IEEE 2011 workshop on automatic speech recognition and understanding**. [S.l.], 2011. Citations on pages 50, 56, 72, 74, 75, 77, and 101.

QUINTANILHA, I. M. **End-to-End Speech Recognition Applied to Brazilian Portuguese Using Deep Learning**. Phd Thesis (PhD Thesis) — MSc dissertation, PEE/-COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, 2017. Citations on pages 17, 44, and 45.

RAAB, M.; GRUHN, R.; NOETH, E. Non-native speech databases. In: IEEE. **Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on**. [S.l.], 2007. p. 413–418. Citations on pages 25 and 46.

RABINER, L. R.; JUANG, B.-H. **Fundamentals of speech recognition**. [S.l.]: PTR Prentice Hall Englewood Cliffs, 1993. Citation on page 71.

RISTAD, E. S.; YIANILOS, P. N. Learning string edit distance. **IEEE Transactions on Pattern Recognition and Machine Intelligence**, v. 20, n. 5, p. 522–532, May 1998. Citation on page 96.

RYAN, E. B.; CARRANZA, M. A.; MOFFIE, R. W. Reactions toward varying degrees of accentedness in the speech of spanish-english bilinguals. **Language and Speech**, Sage Publications, v. 20, n. 3, p. 267–273, 1977. Citations on pages 28 and 37.

SAHIDULLAH, M.; SAHA, G. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. **Speech Communication**, Elsevier, v. 54, n. 4, p. 543–565, 2012. Citation on page 49.

SAINATH, T. N.; MOHAMED, A.-r.; KINGSBURY, B.; RAMABHADRAN, B. Deep convolutional neural networks for lvcsr. In: IEEE. **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**. [S.l.], 2013. p. 8614–8618. Citations on pages 33, 70, 72, 78, 101, 103, and 117.

SAITO, K.; LYSTER, R. Investigating the pedagogical potential of recasts for l2 vowel acquisition. **TESOL Quarterly**, Wiley Online Library, v. 46, n. 2, p. 387–398, 2012. Citation on page 25.

SAMUDRAVIJAYA, K. Sphinx: a toolkit for asr. 2010. Citations on pages 50 and 74.

SATO, C. J. Phonological processes in second language acquisition: Another look at interlanguage syllable structure. **Language Learning**, Wiley Online Library, v. 34, n. 4, p. 43–58, 1984. Citation on page 40.

SCHIEL, F.; DRAXLER, C.; BAUMANN, A.; ELLBOGEN, T.; STEFFEN, A. The production of speech corpora. **epub uni-muenchen**, 2012. Citation on page 82.

SCHLIPPE, T. Lecture slides. 2012. Accessed: 2017-09-01. Citations on pages 17 and 55.

SCHMIDT, R.; FROTA, S. Developing basic conversational ability in a second language: A case study of an adult learner of portuguese. **Talking to learn: Conversation in second language acquisition**, p. 237–326, 1986. Citations on pages 28, 37, and 38.

SCHUSTER, E.; LEVKOWITZ, H.; OLIVEIRA, O. N. **Writing scientific papers in English successfully: your complete roadmap**. [S.l.]: Hypertek. com, Incorporated, 2014. Citation on page 25.

SELINKER, L.; LAKSHMANAN, U. Language transfer and fossilization: The multiple effects principle. **Language transfer in language learning**, John Benjamins Amsterdam, p. 197–216, 1992. Citations on pages 31 and 36.

SHULBY, C. Teacher preparation for online education. In: IATED. **INTED2013 Proceedings**. [S.l.], 2013. p. 199–207. Citation on page 27.

SHULBY, C. D.; FERREIRA, M. D.; MELLO, R. F. de; ALUÍSIO, S. M. Acoustic modeling using a shallow CNN-HTSVM architecture. In: **2017 Brazilian Conference on Intelligent Systems (BRACIS)**. Piscataway, NJ: IEEE, 2017. p. 85–90. Citations on pages 30 and 118.

_____. Cnn parameter selection and context windows for robust phoneme recognition. In: SPRINGER. **International conference on statistical language and speech processing**. [S.l.], 2018 (Submitted). p. 1–11. Citation on page 121.

SILVEIRA; ROSANE; UBIRATÃ; KICKHÃFEL; ALVES. **Pronunciation instruction for Brazilians: bringing theory and practice together**. [S.l.]: Cambridge Scholars Publishing, 2009. Citation on page 26.

SMITH, D. **Machine Learning with Spreadsheets! Part 2: Dummies Guide to Convolutional Neural Nets**. 2018. <https://medium.com/excel-with-ml/https-medium-com-excelwithml-machine-learning-with-spreadsheets-part-2-convolutional-neural-nets-c67 Accessed: 2018-06-01. Citations on pages 17 and 61.

SOLTAU, H.; SAON, G.; SAINATH, T. N. Joint training of convolutional and non-convolutional neural networks. In: **ICASSP**. [S.l.: s.n.], 2014. p. 5572–5576. Citation on page 72.

SONG, W.; CAI, J. End-to-end deep neural network for automatic speech recognition. Technical Report, 2015. Citations on pages 85 and 143.

SOUZA, E. P. d.; PAULA, M. C. d. S. Qualis: a base de qualificação dos periódicos científicos utilizada na avaliação capes. **InfoCAPES Boletim Informativo**, v. 10, n. 2, 2002. Citation on page 27.

SOUZA, G.; NETO, N. An automatic phonetic aligner for brazilian portuguese with a praat interface. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2016. p. 374–384. Citation on page 83.

STEMBRIDGE, B. **Brazil: A Shining Star Among BRIC Nations**. 2012. Available: <http://www.iam-magazine.com/issues/article.ashx?g=d2cc49a5-4a5d-476a-ab20-4cdfc1fcc850>. Citation on page 26.

STOLCKE, A. Srilm-an extensible language modeling toolkit. In: **Seventh international conference on spoken language processing**. [S.l.: s.n.], 2002. Citation on page 53.

SWAIN, M. The output hypothesis: Theory and research. **Handbook of research in second language teaching and learning**, v. 1, p. 471–483, 2005. Citations on pages 37 and 38.

SWALES, J. The concept of discourse community. **Genre analysis: English in academic and research settings**, p. 21–32, 1990. Citation on page 54.

_____. **Research genres: Explorations and applications**. [S.l.]: Ernst Klett Sprachen, 2004. Citation on page 25.

SWIETOJANSKI, P.; GHOSHAL, A.; RENALS, S. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In: IEEE. **Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on**. [S.l.], 2013. p. 285–290. Citation on page 73.

TAN, T.-P.; BESACIER, L.; LECOUTEUX, B. Acoustic model merging using acoustic models from multilingual speakers for automatic speech recognition. In: IEEE. **Asian Language Processing (IALP), 2014 International Conference on**. [S.l.], 2014. p. 42–45. Citation on page 75.

TAO, J.; GHAFFARZADEGAN, S.; CHEN, L.; ZECHNER, K. Exploring deep learning architectures for automatically grading non-native spontaneous speech. In: IEEE. **Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on**. [S.l.], 2016. p. 6140–6144. Citation on page 75.

THOMSON, R. I.; DERWING, T. M. The effectiveness of l2 pronunciation instruction: A narrative review. **Applied Linguistics**, Oxford University Press, v. 36, n. 3, p. 326–344, 2014. Citations on pages 25 and 27.

TÓTH, L. Phone recognition with hierarchical convolutional deep maxout networks. **EURASIP Journal on Audio, Speech, and Music Processing**, Springer, v. 2015, n. 1, p. 25, 2015. Citation on page 70.

TYAGI, V.; WELLEKENS, C. On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition. In: IEEE. **Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on**. [S.l.], 2005. v. 1, p. I–529. Citation on page 50.

UPADHYAYA, S. R. Parallel approaches to machine learning—a comprehensive survey. **Journal of Parallel and Distributed Computing**, Elsevier, v. 73, n. 3, p. 284–292, 2013. Citations on pages 71 and 73.

VAPNIK, V. **Statistical learning theory**. [S.l.]: Wiley New York, 1998. Citations on pages 33, 63, 64, 65, 113, and 117.

_____. **The nature of statistical learning theory**. [S.l.]: Springer science & business media, 2013. Citations on pages 33, 63, 66, 67, 103, 111, 113, 114, 115, and 135.

VAPNIK, V.; CHERVONENKIS, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. **Theory of Probability and its Applications**, Society for Industrial and Applied Mathematics, v. 16, n. 2, p. 264, 1971. Citation on page 64.

VASILEV, I. A deep learning tutorial: From perceptrons to deep networks. **Toptal**, 2015. Accessed: 2018-03-01. Citations on pages 73 and 74.

VERTANEN, K. **HTK Acoustic Models**. 2018. Accessed: 2018-03-01. Available: <https://www.keithv.com/software/htk/us/>. Citations on pages 101 and 110.

VU, N. T.; WANG, Y.; KLOSE, M.; MIHAYLOVA, Z.; SCHULTZ, T. Improving asr performance on non-native speech using multilingual and crosslingual information. In: **Fifteenth Annual Conference of the International Speech Communication Association**. [S.l.: s.n.], 2014. Citations on pages 25, 29, 77, 88, and 105.

VYGOTSKY, L. S. **Mind in society: The development of higher psychological processes**. [S.l.]: Harvard university press, 1980. Citation on page 38.

WAIBEL, A.; HANAZAWA, T.; HINTON, G.; SHIKANO, K.; LANG, K. J. Phoneme recognition using time-delay neural networks. **IEEE transactions on acoustics, speech, and signal processing**, IEEE, v. 37, n. 3, p. 328–339, 1989. Citation on page 70.

WANG, Z.; SCHULTZ, T.; WAIBEL, A. Comparison of acoustic model adaptation techniques on non-native speech. In: IEEE. **Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on**. [S.l.], 2003. v. 1, p. I–I. Citations on pages 29, 76, 88, and 105.

WEIDE, R. **The Carnegie mellon pronouncing dictionary [cmudict. 0.6]**. [S.l.]: Carnegie Mellon University: http://www. speech. cs. cmu. edu/cgi-bin/cmudict. Accessed, 2005. Citations on pages 56 and 82.

WEINBERGER, S. Speech accent archive. george mason university. **Online:< http://accent. gmu. edu**, 2014. Citations on pages 30 and 38.

WIATOWSKI, T.; BÖLCSKEI, H. A mathematical theory of deep convolutional neural networks for feature extraction. **IEEE Transactions on Information Theory**, IEEE, 2017. Citations on pages 62 and 63.

WITT, S. M. Automatic error detection in pronunciation training: Where we are and where we need to go. In: **International Symposium on Automatic Detection of Errors in Pronunciation Training, Stockholm, Sweden**. [S.l.: s.n.], 2012. Citations on pages 25, 27, and 42.

XU, M.; DUAN, L.-Y.; CAI, J.; CHIA, L.-T.; XU, C.; TIAN, Q. Hmm-based audio keyword generation. In: SPRINGER. **Pacific-Rim Conference on Multimedia**. [S.l.], 2004. p. 566–574. Citation on page 49.

YANG, H. H.; VUUREN, S. V.; SHARMA, S.; HERMANSKY, H. Relevance of time–frequency features for phonetic and speaker-channel classification. **Speech communication**, Elsevier, v. 31, n. 1, p. 35–50, 2000. Citation on page 116.

YOUNG, S.; EVERMANN, G.; GALES, M.; HAIN, T.; KERSHAW, D.; LIU, X.; MOORE, G.; ODELL, J.; OLLASON, D.; POVEY, D. *et al.* The htk book. **Cambridge university engineering department**, v. 3, p. 175, 2002. Citations on pages 49, 50, 74, 82, and 110.

YU, D.; DENG, A. A.; DAHL, G.; SEIDE, F.; LI, G. More data+ deeper model= better accuracy. In: **keynote at International Workshop on Statistical Machine Learning for Speech Processing**. [S.l.: s.n.], 2012. Citation on page 117.

YUAN, J.; LIBERMAN, M. Speaker identification on the scotus corpus. **Journal of the Acoustical Society of America**, [New York: Acoustical Society of America], v. 123, n. 5, p. 3878, 2008. Citations on pages 51, 82, and 97.

ZAREMBA, W.; SUTSKEVER, I.; VINYALS, O. Recurrent neural network regularization. **arXiv preprint arXiv:1409.2329**, 2014. Citation on page 85.

ZHANG, C.; BENGIO, S.; HARDT, M.; RECHT, B.; VINYALS, O. Understanding deep learning requires rethinking generalization. **arXiv preprint arXiv:1611.03530**, 2016. Citations on pages 30, 63, and 73.

ZHANG, Y.; PEZESHKI, M.; BRAKEL, P.; ZHANG, S.; BENGIO, C. L. Y.; COURVILLE, A. Towards end-to-end speech recognition with deep convolutional neural networks. **arXiv preprint arXiv:1701.02720**, 2017. Citations on pages 78 and 83.


ICMC
USP
SÃO CARLOS