Processamento de fala para triagem de distúrbios fonológicos

Guilherme Jun Yoshimura

TEXTO APRESENTADO
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa de Pós-graduação em Ciência da Computação
Orientador: Prof. Dr. Marcelo Gomes de Queiroz

São Paulo, 18 de Maio de 2020

Processamento de fala para triagem de distúrbios fonológicos

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 05/02/2020. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Marcelo Gomes de Queiroz (orientador) - IME-USP
- Profa. Dr. Haydée Fiszbein Wertzner - FM-USP
- Prof. Dr. Tiago Tavares Fernandes - UNICAMP

Agradecimentos

Primeiramente, agradeço ao meu orientador, Marcelo Queiroz, pelo apoio e incentivo no meu trabalho. Desde a escolha do tema do projeto até as discussões de trabalhos futuros, sempre me direcionou de forma precisa. Depois de cada reunião sempre, surgia um novo desafio, uma nova forma de resolver o problema. Foram essas conversas que sempre me motivaram a tentar ir cada vez mais longe. Agradeço pelo seu tempo investido a mim, todo seu conhecimento que você me passou e a paciência que teve em esclarecer minhas dúvidas.

Agradeço aos professores que participaram da banca de qualificação e defesa, Haydée, Miguel e Tiago. Seus argumentos me ajudaram a enriquecer e esclarecer muitos pontos do trabalho.

Agradeço também a Haydée que me acolheu e aceitou minha proposta de pesquisa. Sua paciência em explicar temas que muitas vezes eram básicos, ajudaram muito a entender como eu poderia propor soluções para resolver os problemas do meu mestrado. Além de disso, gostaria de agradecer todo a equipe do Laboratório de Investigação Fonoaudiológica em Fonologia por ter disponibilizado o tempo deles para gerar as bases que utilizei neste trabalho.

Agradeço aos meus professores de graduação Vanderlei Bonato e Roseli Romero que sempre me incentivaram a realizar atividades extracurriculares como iniciação científica, monitorias e intercâmbio. Agradeço por novamente terem incentivado e me apoiado a fazer o mestrado.

Por fim, agradeço aos meus pais Jorge e Rita, meus irmãos Felipe e Alexandre, minha tia Nilce e minha namorada Carolina que sempre estiveram ao meu lado e me deram todo apoio necessário. Sem eles nada disso seria possível.

Resumo

Este trabalho apresenta dois classificadores originais para sinais de voz que objetivam auxiliar profissionais da fonoaudiologia no diagnóstico de pessoas com alterações de fala. Comparamos os classificadores propostos com três técnicas conhecidas: Modelos de Markov Escondidos (HMM), bag-of-words e classificador baseado em Earth Mover's Distance (EMD).

Utilizamos três bases de dados, sendo duas disponibilizadas pelo Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional (FOFITO) da Faculdade de Medicina da Universidade de São Paulo (FMUSP) que contêm gravações de crianças que têm alterações de fala que ocorrem durante o desenvolvimento da fala, e a terceira é a base pública UA-Speech que contêm gravações de indivíduos adultos com disartria.

O intuito deste trabalho é criar classificadores de fala capazes de distinguir um áudio sem alteração de fala de um áudio com alteração de fala. Além de estudar as técnicas conhecidas citadas anteriormente, propusemos dois classificadores baseados em Coeficientes Mel-Cepstrais (MFCC). O primeiro utiliza uma reformulação da distância DTW entre registros de fala e conjuntos de gravações sem alteração de fala, enquanto o outro combina a informação de curvas de dissimilaridades construídas a partir da comparação do registro de fala a ser classificado com as gravações de referência (sem alterações de fala).

Palavras-chaves: Processamento de fala, Distúrbio do som da fala, Dynamic Time Warping, classificação, Coeficientes Mel-Cepstrais.

Abstract

This work presents two novel speech classifiers which aim to aid speech therapy professionals in the diagnosis of individuals with speech disorders. We compared the proposed classifiers with three well-known techniques: Hidden Markov Models (HMM), Bag-of-Words (BoW) and a classifier based on the Earth Mover's Distance.

In this work we used three databases, two of which were provided by the School of Medicine at the University of São Paulo, and a third one which is a public database (UA-Speech) containing recordings of individuals with dysarthria.

The goal of this project is to develop speech classifiers which are able to distinguish recordings from patients with and without speech disturbances. Besides studying the well-known techniques mentioned above, we proposed two techniques that are based on Mel Frequency Cepstral Coefficients (MFCC). The first one defines the classification problem over relative embeddings based on point-to-set distances, while the second one combines information from dissimilarity curves built from the comparison of the speech recording to be classified and the reference recordings (without speech disorders).

Keywords: Speech Processing, Speech Sound Disorder, Dynamic Time Warping, Classification, Mel Frequency Cepstral Coefficients.

Sumário

Lista de Abreviaturas	vi
Lista de Figuras	vii
1 Introdução	1
1.1 Considerações Preliminares	1
1.1.1 Alterações de fala	2
1.1.2 Motivações	3
1.1.3 Aplicações	3
1.2 Visão geral de um classificador de fala	5
1.3 Desafios	6
1.4 Objetivos	7
1.5 Organização do trabalho	7
2 Ferramentas e métodos	9
2.1 Base de dados	9
2.1.1 Base de dados da Fonoaudiologia da USP	9
2.1.2 UA Speech Database	9
2.2 Ferramentas	11
2.2.1 Librosa	11
2.2.2 Scikit.Learn	11
2.2.3 hmmlearn	11
2.2.4 Scipy	11
2.3 Características	11
2.3.1 MFCC	12
2.3.2 Bag of Words (BoW)	13
2.4 Métodos de classificação	13
2.4.1 Modelos Ocultos de Markov(HMM)	13
2.5 Cálculo de similaridade	14
2.5.1 DTW	14
2.5.2 Earth mover's distance	15
2.6 Métodos de validação de algoritmos	16
2.6.1 Validação Cruzada	16
2.6.2 F-measure e Acurácia	16
2.6.3 F-measure com validação cruzada	17

3	Desenvolvimento e Metodologia	19
3.1	Desenvolvimento	19
3.1.1	Imersão por distância DTW relativa	19
3.1.2	Classificador baseado em perfis de dissimilaridade realinhados	22
3.1.3	Vector Quantization + Bag-of-Words	23
3.1.4	Modelo Oculto de Markov	23
3.1.5	Classificador baseado na <i>Earth movers distance</i>	24
3.2	Metodologia Experimental	25
3.2.1	Metodologia experimental MinDTW	25
3.2.2	Metodologia experimental TADPC	26
3.2.3	Metodologia experimental Vector Quantization + Bag-of-Words	26
3.2.4	Metodologia experimental Modelo de Markov oculto	27
3.2.5	Classificador baseado na <i>Earth movers distance</i>	27
4	Resultados experimentais	29
4.1	Resultados utilizando as bases de dados do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP	29
4.2	Resultados utilizando a base de dados UA-Speech	34
4.3	Discussão	38
5	Considerações finais	41
5.1	Conclusão	41
5.2	Trabalhos futuros	41
	Referências Bibliográficas	44

Lista de Abreviaturas

MFCC	Coeficientes Mel-Cepstrais (<i>Mel-frequency cepstral coefficients</i>)
DTW	<i>Dynamic Time Warping</i>
HMM	Modelo de Markov escondido (<i>Hidden Markov Model</i>)
BoW	<i>Bag-of-Words</i>
VQ	Quantização vetorial(<i>Vector Quantization</i>)
ANN	Redes Neurais Aritificias (<i>Artificial Neural Network</i>)
EMD	<i>Earth Mover's Distance</i>
TADPC	Classificador baseado em perfis de dissimilaridade realinhados (<i>Time-aligned Dissimilarity Profile Classifier</i>)

Lista de Figuras

1.1	Prova de imitação aplicada durante a triagem.	4
1.2	Prova de nomeação aplicada durante a triagem.	4
1.3	Exemplo de uma imagem da prova de nomeação.	5
2.1	Equipamento utilizado para coleta dos áudios e vídeos da base UA-Speech [KHJP ⁺ 08]	11
2.2	Diagrama de blocos para obtenção do MFCC	12
2.3	Banco de filtros na escala Mel	13
2.4	Imagem retirada do livro [Mül15] (a) Matriz de custo da sequência $X = (1, 3, 3, 8, 1)$ e $Y = (2, 0, 0, 8, 7, 2)$. (b) Matriz de custo acumulado e caminho ótimo. (c) Resultado do alinhamento.	15
2.5	Comparação entre os cálculos de F measure utilizando validação cruzada [FS10] . . .	18
3.1	Classificação binária entre itens heterogêneos. Itens podem ser MFCCgramas ou Cromagramas, ou qualquer outro tipo de matriz de características obtidas quadro- a-quadro.	20
3.2	Classificação binária utilizando o método de imersão por distância DTW re- lativa . O item x é associado à sua distância até o representante mais próximo da classe $\mathcal{N}$	20
3.3	Etapas do classificador baseado em perfis de dissimilaridade realinhados	22
3.4	Comparação da curva de um áudio sem alteração de fala e de um áudio com alteração de fala	23
3.5	funções de densidade de probabilidade referentes à distribuição de todos os áudios sem alteração de fala (verde), a uma gravação sem alteração de fala específica (azul) e a uma gravação com alteração de fala (laranja), utilizadas no classificador baseado na <i>Earth movers distance</i>	24
3.6	Matriz de similaridade da palavra “Seven” na base de dados UA-Speech. $\mathcal{N} =$ $\{0, \dots, 39\}$ corresponde à classe dos áudios rotulados como sem alteração de fala e o restante é a classe dos áudios que apresentam algum alteração de fala.	25
3.7	Resultado do alinhamento de um áudio com todos os áudios sem alteração de fala da base de treinamento	26

4.1	Classificação da palavra “Chuva” usando <i>imersão MinDTW relativa</i> . Círculos e quadrados representam os dados de treinamento e triângulos os dados de teste; o eixo horizontal é referente à distância relativa até a classe sem alteração de fala; linhas verticais representam (da esquerda para direita) ϱ^+ , τ e ϱ^-	30
4.2	Classificação da palavra “Chuva” usando <i>imersão HausdorffDTW relativa</i> . As convenções utilizadas são as mesmas da Figura 4.1.	30
4.3	Classificação da palavra “Chuva” usando VQ+BoW.	31
4.4	Classificação da palavra “Chuva” usando HMM.	31
4.5	Valor da F-measure global dos métodos MinDTW, HausdorffDTW, VQ+BoW e HMM	31
4.6	Todas as curvas de dissimilaridade criadas a partir de uma gravação com alteração de fala da palavra “chave” comparada com as gravações sem alteração de fala da base de treino.	32
4.7	Mesma análise da Figura 4.6 utilizando outra gravação com alteração de fala da palavra “chave”.	32
4.8	Todas as curvas de dissimilaridade criadas a partir de uma gravação com alteração de fala da palavra “sapo” comparada com as gravações sem alteração de fala da base de treino.	33
4.9	Mesma análise da Figura 4.8 utilizando outra gravação com alteração de fala da palavra “sapo”.	33
4.10	Classificação da palavra “Command” usando <i>imersão DTW relativa</i>	34
4.11	Classificação da palavra “Hypothesis” usando <i>imersão DTW relativa</i>	34
4.12	Classificação da palavra “Command” utilizando VQ+BoW; o eixo horizontal é a distância até o centroide da classe de palavras sem alteração de fala.	35
4.13	Classificação da palavra “Hypothesis” utilizando VQ+BoW.	35
4.14	Classificação da palavra “Command” usando HMM; o eixo horizontal é o valor absoluto do logaritmo da probabilidade de Viterbi.	35
4.15	Classificação da palavra “Hypothesis” usando HMM.	36
4.16	Valor da F-measure global utilizando a base de dados UA-Speech original.	36
4.17	Valor da F-measure global para os dados da base UA-Speech estendida.	37
4.18	Valor da F-measure global para os dados da base UA-Speech.	38
4.19	Valor da F-measure global para os dados da base UA-Speech.	39
5.1	Matriz com os valores área dos segmentos dos áudios	42
5.2	Resultado da segmentação utilizando TADPS.	42
5.3	Análise do áudio utilizando o software praet, a esquerda temos a representação do áudio inteiro e a direita do áudio segmentado.	43

Capítulo 1

Introdução

A fala é um dos meios mais utilizados para a comunicação humana. Por meio dela, é possível expressar sentimentos, ideias e pensamentos. Nesse sentido, problemas fonológicos podem ser encontrados durante o desenvolvimento da fala, atingindo em média 5% da população pré-escolar [NBMM06], ou decorrente de algum tipo de acidente vascular por exemplo, a disartria que atinge cerca de 280 pessoas a cada 100.000 [EE96]

Muitas alterações de fala que ocorrem na infância poderiam ser evitados ou minimizados com medidas simples, como por exemplo a identificação precoce [PdOM11]. Quando não tratados, podem ter impacto direto no decorrer da vida social, acadêmica e profissional do indivíduo. Dentre os impactos causados, podemos citar o atraso ou o desenvolvimento atípico relacionados a audição, fala e/ou linguagem, em níveis variados de gravidade [NBMM06].

Outras possíveis causas de alterações de fala podem decorrer de algum problema de saúde que o indivíduo teve durante a infância, como por exemplo, o número de vezes que a criança teve otite média sendo que, nesse caso, essa característica influencia a percepção de diferenças na duração entre os sons fricativos surdos e sonoros [WER09]. Além disso, existem também as pessoas que adquirem alterações de fala após algum trauma, tal como a disartria, que é uma dificuldade motora de fala e portanto a alteração apresentada pelo indivíduo é fonética ou seja, há dificuldade em executar os movimentos necessários para produzir um determinado som.

Normalmente os pais são os primeiros a detectar o problema na fala da criança, por exemplo pela dificuldade em compreender o que a criança fala. Quando detectado, a criança deve ser encaminhada para um fonoaudiólogo que aplicará um teste de triagem para detectar o nível do problema e, então propor um tratamento adequado. Atualmente, os testes de triagem aplicados nos pacientes são analisados de forma manual ou seja, o fonoaudiólogo transcreve o teste aplicado que é gravado em áudio e vídeo, para então realizar a avaliação do paciente. No entanto, esse processo manual pode ser subjetivo, oneroso e passível de algum erro.

A fim de tornar este processo de triagem mais rápido e consistente, nas últimas décadas pesquisadores têm desenvolvido métodos de classificação de fala, dentre os quais podemos destacar técnicas de classificação de fala baseada em Coeficientes Mel-Cepstrais (MFCC), combinadas com algoritmos de classificação como cadeias de Markov escondidas (HMM) [JNM16] ou máquinas de vetores de suporte (SVM) [RRN09].

1.1 Considerações Preliminares

Nesta seção serão detalhadas as características fundamentais das alterações de fala; na Seção 1.1.2 serão abordadas as vantagens na automatização do processo de triagem de pacientes e por fim, na seção 1.1.3 serão apresentadas os possíveis contextos de aplicação.

1.1.1 Alterações de fala

Alterações de fala são caracterizadas pela dificuldade ou impossibilidade de uma pessoa (criança ou adulto) produzir corretamente o som de uma palavra. As alterações de fala podem estar presentes em pessoas de qualquer idade e sexo e podem ter diversas causas. Neste trabalho vamos considerar dois tipos de alterações: a disartria em indivíduos adultos e alterações que ocorrem durante o desenvolvimento da fala na infância.

A disartria é o enfraquecimento dos músculos responsáveis pela fala (músculos da face, língua, lábios e garganta, e músculos responsáveis pela respiração) decorrente de um dano cerebral, sendo que esse tipo de distúrbio é conhecido como distúrbio motor de fala.

Qualquer tipo de dano cerebral pode causar a disartria, como por exemplo os danos associados à doença de Parkinson, à esclerose múltipla, a acidentes cardiovasculares e problemas congênitos, etc. Nesse caso, as alterações da fala de uma pessoa com disartria podem incluir uma velocidade da fala muito rápida ou lenta, a incapacidade de movimentar a língua, voz baixa, etc [ASH18].

Outros tipos de alterações podem acontecer na fase de desenvolvimento da fala, que podem estar associados à produção inadequada dos sons, bem como ao uso inadequado das regras fonológicas da língua quanto à distribuição dos sons e tipo de estrutura silábica. Suas causas são desconhecidas, e o grau de gravidade e inteligibilidade de fala são variados [WER02].

As alterações de fala em crianças podem ter várias causas, que podem ser classificadas em 5 subtipos: atraso de fala genético; otite média com efusão; apraxia de desenvolvimento; envolvimento psicológico; e erros residuais. Além disso, aspectos relacionados a gênero e idade também exercem influência [SHR01][SHR02][SLD03][WER06].

Os três testes mais comuns para diagnosticar se uma criança possui alguma alteração de fala são: imitação, nomeação e fala espontânea. Cada uma tem uma abordagem diferente quanto à estimulação do paciente [Wer03]:

- **Fala Espontânea:** existem duas formas de aplicar este teste: direta ou indireta. Na forma direta, são utilizadas questões para as crianças responderem com a nomeação de algum objeto com que a criança esteja familiarizada. A forma indireta pode ser feita por meio de tópicos ou histórias escolhidas pela criança [Shr85].
- **Provas de Imitação:** nesta prova o paciente precisa repetir uma frase ou vocábulo. Para cada idioma existe uma prova diferente, por exemplo para a língua portuguesa existe o Teste de Linguagem Infantil ABFW [WHF04] e para língua inglesa existe o Goldman-Fristoe [Fri86].
- **Provas de Nomeação:** o fonoaudiólogo apresenta uma figura ou objeto ao paciente e este nomeia o objeto. Caso o sujeito não reconheça o objeto, o fonoaudiólogo pode dizer o nome e voltar a pedir a nomeação após mostrar outros itens.

O processo de triagem aplicado no Laboratório de investigação fonoaudiológica em Fonologia do curso de Fonoaudiologia FMUSP, é baseado na prova fonológica ABFW, que tem como propósito avaliar 14 processos fonológicos. São eles:

- **Redução de sílaba:** quando há perda de uma das sílabas do vocábulo;
- **Harmonia consonantal:** quando um fonema sofre interferência de outro fonema vizinho que o antecede ou o segue;
- **Plosivação de fricativas:** o modo de articulação dos fonemas fricativos é transformado em um fonema plosivo;
- **Posteriorização para velar:** um fonema plosivo linguodental se transforma em um plosivo velar;
- **Posteriorização para palatal:** quando há alteração na zona de articulação transformando um fonema fricativo palatal em um fonema fricativo alveolar;

- **Frontalização de velar:** quando há anteriorização de um fonema velar para um fonema plosivo linguo-alveolar;
- **Frontalização de palatal:** quando anteoriza a produção de um fonema fricativo palatal;
- **Simplificação de líquida:** quando há substituição, semi-vocalização e a omissão das vibrantes;
- **Simplificação da consoante final:** quando elimina-se um dos membros do encontro consonantal (ClV e CrV);
- **Simplificação de encontro consonantal:** quando se elimina ou substitui a consoante final do vocábulo ou da sílaba;
- **Sonorização de plosiva:** quando um fonema plosivo surdo é substituído pelo correspondente sonoro;
- **Sonorização de fricativa:** quando um fonema fricativo surdo é substituído pelo correspondente sonoro
- **Ensurdecimento de fricativa:** quando um fonema fricativo sonoro é substituído pelo correspondente surdo.

As provas de fonologia do teste ABFW são divididas em duas: de imitação e de nomeação. A prova de imitação é formada por 39 palavras pré-definidas que são faladas para o paciente e ele precisa repetir a palavra, uma a uma. Na prova de nomeação são apresentadas imagens de palavras pré-definidas, diferentes das palavras apresentadas na prova de imitação, e o paciente precisa nomear as imagens. As Figuras 1.1 e 1.2, mostram as provas de imitação e repetição, respectivamente, e a Figura 1.3, é um exemplo de uma imagem que o paciente precisa nomear.

Toda avaliação é gravada e filmada, e depois é analisada e revisada por profissionais da área. Nesta etapa, o fonoaudiólogo avalia se a criança possui ou não algum tipo de alteração de fala e qual a criticidade desta alteração.

Neste trabalho, vamos considerar a classificação de palavras isoladas, como as utilizadas nas provas de imitação e nomeação. Em particular, utilizaremos duas bases de dados obtidas através do laboratório de Investigação Fonoaudiológica em Fonologia, do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP, bem como a base de dados *UA-Speech Database*. Os experimentos consideram a classificação de palavras isoladas, de forma independente, ou seja teremos um resultado para cada palavra de cada um dos banco de dados.

1.1.2 Motivações

O objetivo deste trabalho, é auxiliar o processo de triagem, criando um classificador que irá utilizar a base de dados fornecida pelo grupo do laboratório de Investigação Fonoaudiológica em Fonologia, do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP, para classificar se o paciente possui ou não alteração de fala. Desta forma, será possível ter um resultado padronizado de forma rápida, baseado em gravações históricas, sendo necessário o acompanhamento de um fonoaudiólogo para validar o resultado da classificação feito pelo algoritmo.

Outro fator motivacional para a realização deste projeto é a falta de estudos de classificação de fala na língua portuguesa, uma vez que a base de dados de pessoas com alterações de fala é bastante escassa.

1.1.3 Aplicações

Por mais que a motivação deste trabalho seja realizar a classificação de fala em áudios em português brasileiro, as técnicas estudadas e propostas poderiam ser usadas em outros bancos de dados e possivelmente em outros contextos, além do distúrbio de fala. No Capítulo 4, mostramos

ARFV - TESTE DE LINGUAGEM INFANTIL
NAS ÁREAS DE FONOLOGIA, VOCABULÁRIO, FLUÊNCIA E PRAGMÁTICA

CAPÍTULO 1 - FONOLOGIA
HAYDÉE FISZBEIN WERTZNER

ANEXO 1
FONOLOGIA. PROTOCOLO DE REGISTRO - IMITAÇÃO

Nome: _____	Acerto: _____	Distorção: _____
Data do Exame: _____	Omissão: _____	
Idade: _____	Substituição: _____	

Vocabulário	Transcrição	Fonema	Inicial	Final
01. Peteca		p		
02. Bandeja		b		
03. Tigela		t		
04. Doce		d		
05. Cortina		k		
06. Gato		g		
07. Foguete		f		
08. Vinho		v		
09. Selo		s		
10. Zero		z		
11. Chuva		ʃ		
12. Jacaré		ʒ		
13. Machado		m		
14. Nata		n		
15. Lama		ɲ		
16. Ônibus		l		
17. Prego		ʎ		
18. Café		ɛ		
19. Alface		r		
20. Raposa		pR		
21. Borracha		bR		
22. Abelha		tR		
23. Carro		dR		
24. Branco		kR		

Figura 1.1: Prova de imitação aplicada durante a triagem.

ARFV - TESTE DE LINGUAGEM INFANTIL
NAS ÁREAS DE FONOLOGIA, VOCABULÁRIO, FLUÊNCIA E PRAGMÁTICA

CAPÍTULO 1 - FONOLOGIA
HAYDÉE FISZBEIN WERTZNER

ANEXO 3
FONOLOGIA. ANÁLISE DOS PROCESSOS FONOLÓGICOS - NOMEAÇÃO

Nome: _____															
Idade: _____	Data: _____														

	palhaço	bolta	tesoura	cadeira	galinha	vestiura	cabota	xiara	mena	navio	livro	sapo	tambor	sapato	balde	faca	fogão	total
Transcrição																		
redução de sílaba																		
harmonia consonantal																		
plosivação de fricativas																		
posteriorização para velar																		
posteriorização para palatal																		
frontalização de velares																		
frontalização de palatal																		
simplificação de líquida																		
simplificação do encontro consonantal																		
simplificação da consoante final																		
sonorização de plosivas																		
sonorização de fricativas																		
ensurdecimento de plosivas																		
ensurdecimento de fricativas																		
outros																		
Total																		

Legenda: processos fonológicos observados durante o desenvolvimento;
processos fonológicos não observados frequentemente durante o desenvolvimento.

Figura 1.2: Prova de nomeação aplicada durante a triagem.

Figura 1.3: Exemplo de uma imagem da prova de nomeação.

resultados de experimentos preliminares utilizando tanto os bancos de dados do Laboratório de Investigação Fonoaudiológica em Fonologia, do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP, quanto a base de dados *UA Speech Database*, que possui dados de pacientes com disartria decorrente de um acidente vascular.

1.2 Visão geral de um classificador de fala

Nas últimas décadas, vários cientistas vêm desenvolvendo métodos de processamento de sinais digitais (*DSP*) e inteligência artificial (*AI*) para classificar pessoas com alteração de fala (Tabela 1.1). Podemos separar o processo de classificação de fala em três etapas: escolha da base de dados; extração das características do áudio; e classificação propriamente dita. As bases de dados e os algoritmos utilizados neste trabalho estão descritos no Capítulo 2.

Base de dados

Uma base de dados ideal é gravada em ambiente onde não há interferências (vozes de outras pessoas, ruídos) e é realizada com equipamentos específicos para este propósito (microfones, gravadores e câmeras de vídeos). Além disso, todos os arquivos de áudios idealmente devem ser gravados no mesmo ambiente e nas mesmas condições, e utilizando os mesmos equipamentos, a fim de evitar diferenças nas gravações que prejudiquem sua comparação.

A escolha da base de dados impacta diretamente no processo de classificação de fala. Se a base de dados for gravada num ambiente não controlado, ou seja, sem os equipamentos adequados, num ambiente não isolado e sem um padrão para gravação, o resultado do algoritmo de classificação de fala pode ser afetado. Nesse caso, pode ser necessário realizar algum tipo de pré-processamento do áudio a fim de atenuar os problemas decorrentes do registro inicial.

Bases de dados contendo uma palavra por arquivo de áudio podem ser utilizadas para detectar alterações do tipo troca de fonemas, prolongação de fonema, repetição e palavras incompletas. Arquivos de áudio que contêm frases inteiras podem ser utilizados por exemplo, para detectar pausas incomuns entre palavras e interjeições.

Características Uma característica de áudio é qualquer aspecto qualitativo ou quantitativo mensurável do som [Bul07]. Ao dizermos que um som é alto ou baixo, estamos utilizando nosso ouvido para captar o som e enviar para o nosso cérebro processar o áudio e extrair dele, características perceptualmente significativas. A metodologia padrão de classificação de áudio, envolve a extração de características de áudio, usadas como representação alternativa ao áudio bruto no algoritmo de classificação. Dentre as características mais usadas no contexto de alterações de fala podemos citar os Coeficientes Mel-Cepstrais (MFCC), a Linear Predictive Coding (LPC) e os Linear Prediction Cepstral Coefficients (LPCC).

Classificadores

Os algoritmos de classificação podem ser supervisionados ou não supervisionados. Os algoritmos supervisionados, utilizam uma base previamente rotulada, com exemplos similares aos dados que serão classificados, para que o algoritmo possa ajustar seus parâmetros a partir dos exemplos de

Artigo	Característica	Classificador	Base de dados	Resultado	Ano
[RRN09]	MFCC	SVM	15 adultos	98.35%	2009
[THLA ⁺ 07]	MFCC	HMM	20 amostras de áudios sem alterações e 15 amostras com alterações de fala	96%	2007
[WKJSS07a]	MFCC	HMM	24 áudios com alterações de fala (10 com distúrbios de repetição e 14 com prolongação de sons fricativos)	70%	2007
[CAHY09a]	LPCC	LDA, KNN	UCLASS	89,77%	2009
[KMRN08]	MFCC	ANN	gravação de 10 pessoas falando 150 palavras	83%	2008
[NNH ⁺ 00]	número de alterações encontradas no áudio e duração do discurso	HMM	69 gravações	99%	2000
[WKJSS07b]	MFCC	HMM	5 fragmentos de cada fonema (s, z, x, v, f)	82%	2007
[GPAKR00]		ANN	51 crianças	92%	2000
[HCAY12]	LPC, LPCC, WLPCC	KNN,LDA	UCLASS	97,45%	2012
[CAMYSC12]	MFCC ,LPCC	KNN,LDA	UCLASS	92,75%	2012
[CAHY09b]	MFCC	KNN, LDA	UCLASS	98,18%	2009
[Mah12]	MFCC	VQ Codebook	UCLASS	100%	2012
[CCF ⁺ 12]	PLP	HMM	UA-Speech	90.2%	2012

Tabela 1.1: *Artigos de classificação de pessoas com alterações de fala*

modo a permitir a classificação de dados novos. Podemos citar como algoritmos de classificação supervisionado o K vizinhos mais próximos (KNN) [PCY⁺07], o algoritmo ingênuo de Bayes (Naive Bayes) [PJS13] e as máquinas de vetor-suporte (SVM) [GHP04].

Classificadores não supervisionados, realizam a classificação sem nenhuma informação prévia dos dados que serão classificados, através de algoritmos de aprendizado de máquina que percorrem a base de dados construindo modelos baseados em alguma medida de similaridade entre os dados a fim de permitir algum tipo de separação. Alguns classificadores não supervisionados conhecidos são o K-means e o algoritmo Expectation–Maximization (EM).

1.3 Desafios

Na área de pesquisa em classificação de fala existem diversos problemas relevantes e não solucionados, sendo que alguns foram identificados em artigos, e outros encontrados durante a pesquisa. Primeiro, vamos descrever os problemas encontrados na literatura, e depois vamos expor algumas dificuldades que encontramos durante a pesquisa.

- **Base de dados:** Encontrar uma base de dados pública e padronizada com um número considerável de áudios de pessoas com alterações de fala é muito difícil. Em língua inglesa, muitos

autores utilizam bases pequenas, o que acaba limitando a abrangência do resultado obtido no trabalho. Podemos mencionar duas grandes bases de dados de pessoas com alterações de fala, que são a *UA-Speech Database* criada pelo grupo *Statistical Speech Technology* da universidade de Illinois [KHJP⁺08] e a base de dados *University College London's Archive of Stuttered Speech (UCLASS)* [HP09].

Atualmente, não existe uma base do porte da *UA-Speech Database* ou *UCLASS* na língua portuguesa, mas com a parceria do Laboratório de Investigação Fonoaudiológica em Fono-logia, do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP, conseguimos duas bases de dados. A primeira é uma pequena amostra da base de dados de crianças que possuem alterações de fala e de crianças que não possuem alterações de fala com seus respectivos rótulos ("sem alteração de fala", "com alteração de fala"), e a segunda é uma base de dados contendo mais de 100 gravações.

- **Problemas na qualidade do áudio:** A qualidade do áudio também impacta no processamento da fala. Nas bases de dados que estamos utilizando, os principais problemas são: ruídos e vozes de outros falantes. No caso da base da fonoaudiologia da USP, que contém as gravações do teste ABFW, isso acontece pois o áudio é gravado durante uma consulta; na base de dados UA-Speech também foram detectados alguns casos de mais de uma pessoa falando no áudio.
- **Segmentação do áudio:** Conforme mencionado anteriormente, os registros da base de dados da fonoaudiologia são obtidos durante uma consulta onde é aplicada a prova de imitação e nomeação, exigindo a segmentação prévia do áudio, que hoje é realizada de forma manual. Além de ser um trabalho oneroso, o áudio segmentado pode conter ruídos que impactam na classificação do áudio.
- **Diferenças nos tipos de alterações de fala:** Conforme evoluímos no desenvolvimento de classificadores, foi possível identificar que a precisão do modelo variava de acordo com o tipo de alteração que estávamos lidando. Uma alteração caracterizada por uma fala mais lenta, como na disartria, é bem diferente de um distúrbio caracterizado pela troca de fonemas. Iremos discutir essas diferenças na Seção 4.3.

1.4 Objetivos

Este trabalho tem como principal objetivo estudar classificadores binários que sejam capazes de separar áudios de pessoas com e sem alterações de fala. Para isso, devemos:

- Fundamentar teoricamente o problema de classificação de fala.
- Avaliar os tipos de alterações de fala, considerando as diferenças entre as bases de dados utilizadas.
- Avaliar os modelos de classificação de fala existentes na literatura sobre os bancos de dados utilizados.
- Propor ferramentas e métodos originais para melhorar as métricas de classificação de fala.

1.5 Organização do trabalho

Esta dissertação está estruturada como segue: a fundamentação teórica é exposta no Capítulo 2; alguns métodos existentes de classificação de fala bem como duas propostas novas de classificação são apresentados no Capítulo 3; resultados dos experimentos são apresentados no Capítulo 4; e por fim, uma discussão e trabalhos futuros são apresentados no Capítulo 5.

O Capítulo 2 apresenta as ferramentas utilizadas e métodos estudados para desenvolver o classificador de fala. Neste capítulo, são discutidas as bases de dados (Seção 2.1), as linguagens e bibliotecas de programação (Seção 2.2), e a metodologia, sendo esta dividida em métodos de extração de características de áudio (Seção 2.3), métodos de classificação (Seção 2.4), cálculo de similaridade (Seção 2.5) e métodos de avaliação dos algoritmos (Seção 2.6).

No Capítulo 3 apresentamos como foram realizados os experimentos deste trabalho. Na seção 3.1.1 e 3.1.2 apresentamos duas novas técnicas de classificação de fala baseadas em coeficientes Mel-Cepstrais. A primeira utiliza distância DTW ponto-a-classe, e a segunda utiliza a combinação dos perfis de dissimilaridades obtidos através do DTW de dois sinais de áudio. Por fim, nas seções 3.1.3, 3.1.4 e 3.1.5 apresentamos três abordagens conhecidas Bag-of-Words e modelos de Markov ocultos e classificador baseado em *Earth Mover's Distance* respectivamente.

No Capítulo 4 apresentamos e discutimos os resultados obtidos pelos sistemas apresentados no Capítulo 3, utilizando as bases de dados do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP e UA-Speech. Finalmente, no Capítulo 5 discutimos as dificuldades encontradas durante os experimentos, as contribuições do trabalho e trabalhos futuros. Além disso, discutimos outras abordagens para tentar resolver o problema de classificação de fala, analisando as vantagens e desvantagens em comparação com o método proposto, além de planos futuros para a continuidade do projeto.

Capítulo 2

Ferramentas e métodos

2.1 Base de dados

Para realizar os testes deste projeto, foram utilizadas duas bases disponibilizadas pelo Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP e uma base pública da faculdade de Illinois *UA-Speech Database*.

2.1.1 Base de dados da Fonoaudiologia da USP

A primeira base disponibilizada pelo departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP contém gravações realizadas durante o diagnóstico de crianças com distúrbios associados ao desenvolvimento da fala. Antes de iniciar o tratamento, é realizada a avaliação utilizando os testes de imitação e nomeação mencionados na seção 1.1.1, processo este gravado em áudio e vídeo. Em parceria com o departamento da FMUSP, obtivemos 20 arquivos de áudio de pacientes com idades entre 5 anos e 9 meses até 7 anos e 7 meses, sendo 10 arquivos de áudios de crianças com alterações que ocorrem durante a fase de desenvolvimento da fala.

As gravações foram feitas no consultório do fonoaudiólogo utilizando um microfone conectado ao computador e o software CSL-Kay Pentax. Como todo o teste é gravado e existe uma interação do profissional com o paciente, é necessário realizar uma segmentação do áudio, que hoje é feita de forma manual.

A segunda base fornecida pelo departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP foi criada durante o trabalho da aluna Danira Francisco [FW17] que tinha como foco estudar as alterações que ocorrem na produção dos fonemas [s] e [ʃ].

Esta base é composta por duas palavras em português, chave ['favi] e sapo ['sapu], que são palavras muito utilizadas para o diagnóstico de alterações de fala em crianças. Foram coletados os áudios de 21 crianças, com idades entre 5 e 11 anos, sem histórico familiar ou pessoal de alterações de fala ou lesões auditivas, otológicas ou neurológicas, diagnosticadas ou suspeitas, e sem intervenções fonoaudiológicas anteriores.

Os áudios foram gravados palavra por palavra, em um ambiente controlado. As gravações foram rotuladas de acordo com a pontuação no teste fonológico de linguagem infantil associado à Porcentagem de Consoantes Corretas (ABFW - PCC) [SK82], que é o protocolo oficial utilizado no português do Brasil. Foram feitos 100 registros de cada palavra, sendo 60% rotulados como sem alteração de fala e 40% rotulados como tendo algum tipo de alteração de fala nas consoantes [s] e [ʃ].

2.1.2 UA Speech Database

A base de dados criada pelo grupo Statistical Speech Technology da faculdade de Illinois [KHJP⁺08] é uma base de dados pública, em que o acesso se dá através da solicitação por e-mail a um de seus criadores. A base de dados é composta por gravações de pessoas com alterações de fala, mais especificamente disartria associada a paralisia cerebral.

A base contém gravações para cada indivíduo, de 765 palavras isoladas, sendo elas palavras incomuns, comandos de computador, palavras usadas no alfabeto por rádio, palavras comuns e dígitos; a tabela 2.1 traz as quantidades e alguns exemplos de cada categoria.

Classe de palavras	Exemplos
Números (10 palavras, 3 repetições)	"zero, one,..., nine"
International Radio Alphabet (26 palavras, 3 repetições)	"alpha, bravo, charlie,..."
Comandos de computador (19 palavras, 3 repetições)	"command, line, paragraph, enter,..."
Palavras comuns (100 palavras, 3 repetições)	"the, of, and,..."
Palavras incomuns (300 palavras, 1 repetição)	"naturalization, faithfulness, frugality,..."

Tabela 2.1: Composição da base UA-Speech e exemplos de palavras disponíveis

Além disso, a base dispõe de informações como idade, nível de inteligibilidade e o tipo de disartria de cada indivíduo, conforme mostrado na tabela 2.2.

Indivíduo	Idade	Inteligibilidade da fala	Diagnóstico
M01	>18	muito baixo (10%)	Espástica
M04	>18	muito baixo (2%)	Espástica
M05	21	mid (58%)	Espástica
M06	18	baixo (39%)	Espástica
M07	58	baixo (28%)	Espástica
M08	28	não possui informação	Espástica
M09	18	alto (86%)	Espástica
M10	21	não possui informação	Ambas
M11	48	mid (62%)	Atetótica
M12	19	não possui informação	Ambas
M13	44	não possui informação	Espástica
M14	40	não possui informação	Espástica
F02	30	baixo (29%)	Espástica
F03	51	muito baixo (6%)	Espástica
F04	18	mid (62%)	Atetótica
F05	22	alto (95%)	Espástica
M01	>18	muito baixo (19%)	Espástica
M02	>18	alto (92%)	Espástica
M03	>18	baixo (29%)	Espástica
F01	>18	baixo (19%)	Espástica

Tabela 2.2: Pacientes da base UA-Speech

A Figura 2.1 mostra como os áudios e vídeos foram gravados. Utilizando um monitor para exibir as palavras que os pacientes precisam pronunciar, e os registros são realizados por um microfone de 7 canais colados no topo do monitor do computador e por uma câmera de vídeo.

Figura 2.1: Equipamento utilizado para coleta dos áudios e vídeos da base UA-Speech [KHJP⁺08]

2.2 Ferramentas

Nesta seção descrevemos as ferramentas e bibliotecas computacionais utilizadas para desenvolver este trabalho.

2.2.1 Librosa

Librosa [MRL⁺15] é um pacote Python para análise de áudio. Este pacote possui funções básicas (processamento de áudio, representação espectral, escala de magnitude e conversão em tempo e frequência), funções para extração de características (MFCC, Cromagrama) e para visualização, entre outras.

2.2.2 Scikit.Learn

Scikit Learn [PVG⁺11] é uma biblioteca de aprendizado de máquina para a linguagem Python. É uma biblioteca código aberto, que inclui diversos algoritmos de classificação, regressão e agrupamento, além de interagir com outras bibliotecas do Python, como Numpy e Scipy.

2.2.3 hmmlearn

Hmmlearn [hmm18] é uma biblioteca que possui a implementação do modelo de Markov escondido (HMM) implementado utilizando as bibliotecas scikit-learn, NumPy, SciPy, and matplotlib. Além de ser uma biblioteca de código aberto.

2.2.4 Scipy

Scipy [Sci19] é uma biblioteca para Python de código aberto que contém diversas funções científicas implementadas. Para este projeto estamos utilizando a implementação da distância de Wasserstein conhecida também como *earth mover's distance*.

2.3 Características

A extração de características é um dos pilares no estudo de processamento digital de áudio. É nesta fase que o sinal de áudio é transformado em um vetor de parâmetros representativos, ou seja, os dados brutos são recodificados de forma que apenas as informações mais importantes

para a tarefa em questão estejam disponíveis, tomando o cuidado para que nenhuma informação crítica seja perdida. Estas informações podem ser aplicadas em diversas áreas de estudo, como por exemplo: a mineração de dados, similaridade de áudio, classificação, etc.

As principais características de áudio usadas em classificação de alterações de fala são: LPC (Linear Predictive Coding) [WN76]; PLP (Perceptual Linear Prediction) [Her90]; e MFCC (Coeficientes Mel-Cepstrais). Nesta seção iremos descrever o algoritmo MFCC.

2.3.1 MFCC

A característica MFCC é muito utilizada em problemas de reconhecimento de fala [DM80, THLA⁺07, CAHY09b]. Os MFCCs são tomados a partir do cepstro de um sinal de voz adaptado à escala MEL. Os MFCCs modelam a não-linearidade na percepção de frequência do sistema auditivo humano, utilizando a escala Mel.

A Figura 2.2 mostra as etapas para o cálculo do MFCC, descritos a seguir.

Figura 2.2: Diagrama de blocos para obtenção do MFCC

Pré-enfase: A fase de pré-enfase é utilizada para realçar as frequências altas do espectro. O seguinte filtro FIR é aplicado no sinal de áudio, nesta etapa.

$$y(n) = x(n) - \alpha x(n-1), \quad (2.1)$$

onde $x(n)$ é o sinal de áudio e $0,9 \leq \alpha \leq 1$.

Enquadramento: Nesta fase, o sinal de áudio é dividido em quadros de duração entre 20-39 ms, onde se supõe que há pouca variação de conteúdo espectral. Além disso, existe uma sobreposição de 75% entre dois quadros sucessivos.

Janelamento: A fase de janelamento, serve para minimizar o problema de descontinuidade de borda na análise do sinal de áudio. A janela de Hamming, é muito utilizada devido a sua boa supressão dos lóbulos laterais.

$$x_a = y_a(n) \cdot w(n), \quad (2.2)$$

$$w(n) = \alpha - \beta \cos \frac{2\pi n}{N-1}, \quad (2.3)$$

onde $a = 1, 2, 3, \dots, T$, $\alpha = 0,54$ e $\beta = 0,46$.

FFT: A transformada rápida de Fourier (FFT), é aplicada para obter os espectros de cada quadro.

Filtro Mel: O espectro obtido é representado na escala de frequências mel, sendo separados em faixas de frequência correspondentes a intervalos linearmente espaçados nessa escala (ver a Figura 2.3). A relação entre frequência linear e a escala mel, é dada por:

$$mel(f) = 2595 * \log_{10} \left(1 + \frac{f}{7000} \right). \quad (2.4)$$

Figura 2.3: Banco de filtros na escala Mel

DCT: Esse processo é usado para converter a informação do domínio da frequência para o domínio cepstral. O cepstro é calculado usando a transformada de cosseno discreta (DCT) ou a transformada de Fourier inversa, de onde finalmente são extraídos os MFCCs.

Ao final deste processo temos uma representação compacta do áudio que preserva as principais informações relevantes à classificação dos sinais de fala.

2.3.2 Bag of Words (BoW)

Bag-of-Words (BoW) é uma técnica muito utilizada em classificação de texto, áudio (Bag-of-Frames [1]) e imagens (Bag-of-Visual-Words [PA12]), sendo particularmente útil em problemas onde os vetores de características usados na representação dos dados possuem tamanhos diferentes, pois BoW transforma os vetores de características em um histograma de tamanho fixo.

A primeira etapa do BoW é separar o vetor de características por palavras, ou no caso onde não existem palavras, em clusteres (utilizando por exemplo o algoritmo K-médias). Os centroides resultantes desta etapa são considerados como *codewords*, ou seja o vetor de característica é quantizado. Por fim, é gerado um histograma a partir dos vetores quantizados e com isso é possível representar os dados com vetores de características de mesmo tamanho.

2.4 Métodos de classificação

A etapa de classificação utiliza os dados obtidos na extração de características para classificar o áudio de forma supervisionada ou não supervisionada.

Nesta seção, vamos descrever os algoritmos de classificação que foram utilizados neste trabalho.

2.4.1 Modelos Ocultos de Markov(HMM)

Modelos ocultos de Markov (HMM) são muito utilizados em problemas em que há necessidade de preservar a evolução temporal, ou seja, aqueles em que a ordem dos eventos afetam o resultado final. HMMs são uma extensão do conceito do modelo de Markov, em que as observações são uma função de probabilidade do estado, ou seja, os HMMs são um processo duplamente estocástico, onde um processo não é visível (oculto), mas que pode ser observado através do outro processo estocástico que produz uma sequência de observações [Rab89]. No nosso cenário as observações são os vetores de características do áudio.

Os elementos de um HMM são:

1. N é o número de estados do modelo, denotados por

$$S = \{S_1, S_2, \dots, S_N\}. \quad (2.5)$$

2. M é o número de observações diferentes por estado. Estas observações correspondem ao resultado do modelo. Por exemplo o lançar de uma moeda, M seria igual a cara ou coroa. Denotamos as observações por

$$V = \{v_1, v_2, \dots, v_M\}. \quad (2.6)$$

3. $A = a_{i,j}$ é a matriz de distribuição de probabilidade das transições de estado, onde

$$a_{i,j} = P(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N; \quad (2.7)$$

no caso onde todos os estados estão interligados, então $a_{i,j} > 0, \forall i, j$.

4. B é a distribuição de probabilidade no estado j , $B = b_j(k)$ onde

$$b_j(k) = P(v_k | q_t = S_j), 1 \leq j \leq N, 1 \leq k \leq M. \quad (2.8)$$

5. π é a distribuição inicial, onde

$$\pi_j = P(q_1 = S_j), 1 \leq j \leq N. \quad (2.9)$$

Com esses dados (M, N, A, B, π) o HMM pode ser usado para gerar a sequência de observações

$$O = O_1, O_2, \dots, O_T, \quad (2.10)$$

onde O_t é um símbolo de V e T é o número de observações. De forma resumida podemos denotar um HMM por $\lambda = (A, B, \pi)$.

2.5 Cálculo de similaridade

Nesta seção, apresentamos dois métodos para cálculo de similaridade. O primeiro método é o DTW (Seção 2.5.1) que tem como objetivo alinhar duas séries temporais a fim de calcular sua dissimilaridade, e o segundo método é o *Earth mover's distance* (seção 2.5.2) que tem como objetivo medir o quão similares são duas distribuições de probabilidade.

2.5.1 DTW

Dynamic Time Warp (DTW) foi introduzido por Sakoe e Chiba [SC78] para reconhecimento de palavras pronunciadas, e é um algoritmo de programação dinâmica baseada em normalização do tempo. Esta técnica visa alinhar ou sincronizar duas séries temporais, esticando ou comprimindo-as ao longo do eixo do tempo. As deformações criadas podem ser utilizadas para sincronizar os trechos de maior similaridade entre dois vetores de características variantes no tempo.

Dado dois vetores de características X de tamanho N e Y de tamanho M , o DTW irá buscar qual o melhor caminho que é iniciado no ponto $(0, 0)$ e termina no ponto (N, M) a ser seguido dentro da matriz de custo acumulado.

O objetivo do DTW é compensar diferenças na duração dos áudios, encontrando possíveis alinhamentos entre duas séries, seja pulando certos trechos ou usando o mesmo trecho mais de uma vez. Além disso, o DTW pode ser utilizado para encontrar um alinhamento ótimo entre as duas séries [Mül15], conforme Figura 2.4.

Figura 2.4: Imagem retirada do livro [Mül15] (a) Matriz de custo da sequência $X = (1, 3, 3, 8, 1)$ e $Y = (2, 0, 0, 8, 7, 2)$. (b) Matriz de custo acumulado e caminho ótimo. (c) Resultado do alinhamento.

2.5.2 Earth mover's distance

Earth mover's distance (EMD) é uma métrica muito utilizada em problemas de recuperação de imagens (image retrieval) [RTG00]. Também conhecida como distância de Wasserstein, ela pode ser definida como o menor esforço para transformar uma função de distribuição de probabilidade em outra.

A EMD é definida para duas funções de densidade de probabilidade P e Q representadas como:

$$P = \{(p_1, \omega_{p_1}), \dots, (p_m, \omega_{p_m})\}$$

$$Q = \{(q_1, \omega_{q_1}), \dots, (q_n, \omega_{q_n})\}$$

onde p_i, q_j representam grupos com distribuições simples (por exemplo Gaussianas) e $\omega_{p_i}, \omega_{q_j}$ representam seus respectivos pesos.

A distância entre as distribuições P e Q é definida a partir das distâncias $d_{i,j}$ entre os grupos p_i e q_j através de um problema de programação linear da forma

$$\text{minimizar } \text{CUSTO}(P, Q, F) = \sum_{i=1}^m \sum_{j=1}^n d_{i,j} f_{i,j}, \quad (2.11)$$

respeitando as seguintes condições:

$$f_{i,j} \geq 0 \quad 1 \leq i \leq m, 1 \leq j \leq n \quad (2.12)$$

$$\sum_{i=1}^n f_{i,j} \leq \omega_{p_i} \quad 1 \leq i \leq m \quad (2.13)$$

$$\sum_{j=1}^n f_{i,j} \leq \omega_{q_j} \quad 1 \leq j \leq n \quad (2.14)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{i,j} = \min \left(\sum_{i=1}^m \omega_{p_i}, \sum_{j=1}^n \omega_{q_j} \right), \quad (2.15)$$

onde $f_{i,j}$ são variáveis que representam o peso relativo de cada par (p_i, q_j) na composição da distância final. Esses pesos relativos devem ser não-negativos (equação 2.12), e devem respeitar algumas condições associadas ao transporte de “carga” (a “terra” na analogia que dá nome à distância) entre os grupos das duas distribuições: a soma das cargas transportadas de um grupo p_i fixado para cada q_j não deve ultrapassar o peso ω_{p_i} do grupo p_i (equação 2.13), e analogamente as

cargas transportadas para o grupo q_j não devem ultrapassar o peso ω_{q_j} (equação 2.14); por fim, a equação 2.16 força o transporte da máxima quantidade de carga possível.

A partir da solução ótima f^* do problema de transporte acima, a *Earth mover's distance* é definida como o custo normalizado pelo total das cargas ótimas através da fórmula

$$EMD(P, Q) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{i,j} f_{i,j}^*}{\sum_{i=1}^m \sum_{j=1}^n f_{i,j}^*} \quad (2.16)$$

2.6 Métodos de validação de algoritmos

2.6.1 Validação Cruzada

Validação cruzada (CV) é uma estratégia muito utilizada para avaliação de algoritmos. Como observado por [Lar31], o treinamento de um algoritmo e a avaliação de seu desempenho estatístico nos mesmos dados produzem resultados enviesados. A CV foi criada para corrigir esse problema, a partir da observação de que o teste da saída do algoritmo em novos dados renderia uma boa estimativa de seu desempenho ([MT87], [Sto74], [Gei75]). A ideia por trás da CV é separar os dados conhecidos em duas ou mais partes para estimar a variabilidade no desempenho de cada algoritmo. Parte dos dados são usados para o treino, e o restante é utilizado na avaliação. Dessa forma, pode-se obter uma estatística mais robusta, não só do desempenho médio, mas também da variabilidade deste.

O principal interesse da CV reside na universalidade das heurísticas de divisão de dados. Ela supõe apenas que os dados são distribuídos de forma idêntica, e as amostras de treinamento e validação são independentes.

Portanto, o CV pode ser aplicado a (quase) qualquer algoritmo em (quase) qualquer estrutura, como regressão ([Sto74] [Gei75]), estimativa de densidade ([Rud82], [Sto84]) e classificação ([DW79], [BBL02]) entre muitos outros.

Existem diversas formas como podemos separar os dados de treino e dados de testes. Abaixo, destacamos as mais conhecidas.

- **Leave-one-out (LOO):** método tradicional onde separamos 1 amostra da base de dados para ser utilizada como amostra de teste, enquanto as $n-1$ amostras são utilizadas para treinar o algoritmo. O processo é repetido até que todas as amostras tenham sido utilizadas como amostra de teste. LOO também é conhecido como: delete-one CV [Li87], ou simplesmente CV [Efr83].
- **Leave-p-out [Sha93]:** possui a mesma ideia do LOO, mas ao invés de utilizar apenas uma amostra de teste, são utilizadas p amostras de testes, e $n-p$ amostras são utilizadas para treinar o algoritmo.
- **V-fold CV (VFCV) ou K-fold CV:** introduzido por [Gei75] como uma alternativa para o alto custo que o LOO apresenta, VFCV particiona a base de dados em V grupos de tamanhos iguais a n/V , onde n é o tamanho total de amostras na base de dados, e V é o número de grupos que estamos criando. Desta forma, conseguimos realizar um processo similar ao LOO mas ao invés de separar apenas uma amostra da base de dados, separamos um grupo para teste e o restante para treino. VFCV também é conhecido como k-fold cross validation.

2.6.2 F-measure e Acurácia

Uma forma de calcular a performance de um classificador é utilizando a F-measure, que é derivada de duas outras métricas: precisão e revocação. Precisão e revocação são adequadas para

avaliar problemas de classificação binária, onde o objetivo é encontrar um conjunto de itens (ditos “verdadeiros”) dentro de um conjunto maior (que contém itens “verdadeiros” e “falsos”).

A precisão representa a porção de itens que o sistema encontra e que são verdadeiros. Para obter alta precisão, deve-se descartar qualquer item que pudesse não ser verdadeiro. Revocação descreve a proporção, dentre todos os itens verdadeiros existentes, daqueles itens verdadeiros que um determinado sistema encontra. Quanto maior o número de itens verdadeiros não localizados, menor é o valor de revocação.

Sempre que utilizamos um classificador binário, ele pode retornar um resultado **verdadeiro positivo** (TP) que é quando o classificador marca a amostra que é verdadeira como verdadeira, **verdadeiro negativo** (TN) quando o classificador marca uma amostra que é falsa como falsa, **falso positivo** (FP) quando o classificador marca a amostra que é falsa como verdadeira, e **falso negativo** (FN) quando marca uma amostra que é verdadeira como falsa. A precisão (Pr) e revocação (Re) de um classificador são calculadas com base nos valores de positivos verdadeiros (TP), falsos positivos (FP) e falsos negativos (FN), através das fórmulas de precisão, revocação, F-measure (F) e acurácia (Acc) abaixo:

$$Pr := \frac{TP}{TP + FP} \quad (2.17)$$

$$Re := \frac{TP}{TP + FN} \quad (2.18)$$

$$F := 2 \times \frac{Pr \times Re}{Pr + Re} \quad (2.19)$$

$$Acc := \frac{TP + TN}{TP + TN + FP + FN} \quad (2.20)$$

2.6.3 F-measure com validação cruzada

Quando aplicamos o método de validação cruzada VFCV e queremos medir a F-measure de algum algoritmo, existem 3 formas principais de calcular a F-measure agregada, que iremos chamar de F_{avg} , $F_{pr,re}$ e $F_{tp,tf}$:

- F_{avg} : corresponde a calcular a média simples das F-measures obtidas em cada uma das dobras ($F^{(i)}$).

$$F_{avg} := \frac{1}{k} \times \sum_i^k F^{(i)} \quad (2.21)$$

- $F_{pr,re}$: corresponde a calcular a média das precisões e revocações obtidas em cada uma das dobras e depois calcular a F-measure utilizando a fórmula descrita acima.

$$Pr := \frac{1}{k} \times \sum_i^k Pr^{(i)} \quad (2.22)$$

$$Re := \frac{1}{k} \times \sum_i^k Re^{(i)} \quad (2.23)$$

$$F_{pr,re} := 2 \times (Pr \times Re) / (Pr + Re) \quad (2.24)$$

- $F_{tp,tf}$: corresponde a sumarizar o número de positivos verdadeiros, negativos verdadeiros e negativos falsos em todas as dobras, e depois computar o valor da F-measure seguindo a

fórmula abaixo.

$$TP := \sum_i^k TP^{(i)} \quad (2.25)$$

$$FP := \sum_i^k FP^{(i)} \quad (2.26)$$

$$FN := \sum_i^k FN^{(i)} \quad (2.27)$$

$$F_{tp,tf} := \frac{2 \times TP}{2 \times TP + FP + FN} \quad (2.28)$$

As medidas de F-measure F_{avg} e $F_{pr,re}$ possuem uma variante que consiste em desconsiderar as dobras (folds) que geram valores de precisão igual a 0, que vamos chamar de $\tilde{F}_{avg}$ e $\tilde{F}_{pr,re}$, respectivamente.

De acordo com [FS10], a forma de agregar as dobras da validação cruzada no cálculo da F-measure agregada pode gerar algum resultado viciado dependendo da fórmula utilizada. Dentre as três formas de calcular o F-measure agregada, a que melhor representa a F-measure global do algoritmo, ou seja aquela com o menor viés segundo [FS10], é a terceira ($F_{tp,tf}$), conforme mostra a Figura 2.5.

Figura 2.5: Comparação entre os cálculos de F measure utilizando validação cruzada [FS10]

É possível observar que os métodos F_{avg} e $F_{pr,re}$ geram um viés negativo, uma vez que o F-measure global é penalizado por dobras que possuem precisão igual a zero, pois suas fórmulas 2.21 e 2.22 utilizam esta informação para calcular a F-measure global. Por outro lado, quando desconsideramos as dobras que possuem precisão igual a zero ($\tilde{F}_{avg}$ e $\tilde{F}_{pr,re}$), estamos gerando um viés positivo, pois além de desconsiderar os casos onde a dobra possui 0 positivos verdadeiros (TP) estamos desconsiderando os falsos positivos (FP) e os falsos negativos (FN) encontrados na dobra.

Desta forma, a melhor maneira de calcular a F-measure global é utilizando a fórmula 2.28 que garante que o cálculo final não será tendencioso para mais ou para menos. Adicionalmente, a acurácia 2.20 não é afetada pois ela considera a quantidade de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

Capítulo 3

Desenvolvimento e Metodologia

Neste capítulo, apresentamos duas novas técnicas¹² de classificação de fala baseadas em coeficientes Mel-Cepstrais. A primeira, batizada de MinDTW, utiliza a distancia DTW ponto-a-classe para determinar posições relativas dos áudios em um espaço de distâncias relativas à classe dos áudios sem alteração de fala. A segunda técnica, denominada TADPC, utiliza uma combinação das curvas de dissimilaridade resultantes do alinhamento de um áudio de teste com cada áudio sem alteração de fala da base de treinamento, que são combinadas em um perfil único de dissimilaridade usado para a classificação. Além disso, apresentamos também três abordagens tradicionais para classificação de fala usadas para efeito de comparação: bag-of-words com vetores quantizados (VQ+BoW), modelos de Markov ocultos (HMM) e um classificador baseado na *Earth mover's distance* (EMD). Separamos o capítulo em duas seções, sendo que na primeira explicamos como os métodos foram desenvolvidos e na segunda como os experimentos foram realizados.

3.1 Desenvolvimento

3.1.1 Imersão por distância DTW relativa

Nesta seção, apresentamos uma nova estratégia de classificação de áudios dependente do tempo que estende a distância de DTW entre pares de matrizes de características para distâncias ponto-a-conjunto, que permitem uma imersão relativa de dados multidimensionais heterogêneos (de dimensões diferentes) em um mesmo espaço de distâncias relativas, usado na classificação. Essa imersão é relativa porque os itens não têm uma posição fixa no espaço destino, mas são posicionados apenas em relação ao conjunto de interesse. A motivação é fornecer uma representação substituta da ideia de centroide + raio ilustrada na Figura 3.1, eliminando a necessidade de um centroide e, ainda assim, permitindo a classificação baseada em um simples critério geométrico.

Considere um conjunto de itens ou matrizes de características $\mathcal{N} = \{M^0, M^1, \dots, M^N\}$ que compõem a classe de interesse para o problema de classificação. Vamos estender a distância DTW regular entre os itens para permitir o cálculo das distâncias ponto-a-classe para qualquer item dentro deste espaço de características heterogêneo. Especificamente, seja x um item arbitrário (dentro ou fora da classe de interesse) do espaço de características, e seja

$$\text{MinDTW}(x) = \min_{y \in \mathcal{N} \setminus \{x\}} \text{DTW}(x, y), \quad (3.1)$$

a menor distância DTW de x para qualquer (outro) item y dentro da classe, isto é, $\text{MinDTW}(x)$

---

¹O material dos capítulos 3.1.1, 4.1 e 4.2 correspondem ao artigo “Relative DTW embedding for binary classification of audio data” de Marcelo Queiroz e Guilherme Jun Yoshimura, apresentado e publicado nos Proceedings of the Sound and Music Computing Conference, Chipre, 2018.

²O material dos capítulos 3.1.2, 4.1 e 4.2 correspondem ao artigo “Speech Sound Disorder Classification Based on Time-Aligned Dissimilarity Profiles” de Guilherme Jun Yoshimura, Marcelo Queiroz, Haydée Fiszbein Wertzner e Danira Francisco, apresentado e publicado nos Proceedings of the Sound and Music Computing Conference, Turim, 2020.

Figura 3.1: Classificação binária entre itens heterogêneos. Itens podem ser MFCCgramas ou Cromagramas, ou qualquer outro tipo de matriz de características obtidas quadro-a-quadro.

expressa quão perto x está do representante mais próximo da classe $\mathcal{N}$ que não seja ele mesmo. O mapeamento $x \mapsto \text{MinDTW}(x)$ é denominado **imersão por distância DTW relativa** de x , uma vez que não posiciona x em um espaço de representação absoluto, mas o posiciona relativamente à classe $\mathcal{N}$.

Tomemos por exemplo a Figura 3.2, onde uma classe $\mathcal{N}$ consiste em 3 itens azuis, e há 2 itens vermelhos fora da classe. Para cada item x , $\text{MinDTW}(x)$ é representado por uma seta de saída partindo de x e alcançando o representante $y \in \mathcal{N} \setminus \{x\}$ mais próximo de x . Nesse caso, os dois itens inferiores da classe são muito próximos um do outro e seus valores $\text{MinDTW}(x)$ são os mesmos; a borda da classe fica relativamente mais distante, e itens fora da classe podem ser acessados por um caminho mais longo. Deve-se notar que as distâncias entre os itens fora da classe não são usadas na imersão, e também quando novos itens são incluídos na classe, todos os valores $\text{MinDTW}(x)$ diminuem ou permanecem os mesmos (por monotonicidade do operador min em relação à inclusão de conjuntos).

Figura 3.2: Classificação binária utilizando o método de **imersão por distância DTW relativa**. O item x é associado à sua distância até o representante mais próximo da classe $\mathcal{N}$.

A **imersão por distância DTW relativa** é definida para todos os itens em um espaço de características heterogêneo, e sua utilidade depende de uma certa propriedade da classe $\mathcal{N}$: que os itens da classe sejam próximos uns dos outros em relação aos itens fora de classe. Em outras palavras, supõe-se, para fins de aplicabilidade deste modelo, que as distâncias intra-classe $\text{DTW}(x, y)$ para $x, y \in \mathcal{N}$ sejam geralmente menores que as distâncias $\text{DTW}(x, y)$ entre itens dentro e fora da classe ($x \in \mathcal{N}, y \notin \mathcal{N}$), ou equivalentemente, que as distribuições estatísticas das

distâncias intra-classe e inter-classes³ sejam significativamente diferentes.

Com base no pressuposto acima, uma estratégia de classificação simples pode ser definida por meio da caracterização das bordas dos itens dentro e fora da classe. Especificamente, sejam

$$\varrho^+ = \max_{z \in \mathcal{N}} \text{MinDTW}(z) \quad (3.2)$$

e

$$\varrho^- = \min_{z \notin \mathcal{N}} \text{MinDTW}(z) \quad (3.3)$$

a maior distância intra-classe e a menor distância inter-classes. Se acontecer que $\varrho^+ < \varrho^-$, uma separação perfeita entre itens dentro e fora da classe é possível, e um limiar intermediário, como

$$\tau = \frac{\varrho^+ + \varrho^-}{2} \quad (3.4)$$

pode ser usado para classificação de novos itens desconhecidos, de acordo com

$$\begin{cases} z \in \mathcal{N} & \text{se } \text{MinDTW}(z) < \tau \\ z \notin \mathcal{N} & \text{caso contrário.} \end{cases} \quad (3.5)$$

Em geral, pode acontecer que os conjuntos das imersões DTW relativas dos elementos internos e externos à classe não sejam perfeitamente separáveis (ou seja, $\varrho^+ \geq \varrho^-$), e então um limiar mais adequado pode ser definido tomando-se o valor ótimo $\tau \in [\varrho^-, \varrho^+]$ de acordo com alguma função objetivo, por exemplo

$$\tau = \underset{\alpha \in [\varrho^-, \varrho^+]}{\text{argmax}} \text{F-measure}(\alpha), \quad (3.6)$$

onde a F-measure é calculada aplicando a estratégia de classificação acima a todos os itens rotulados conhecidos disponíveis para serem usados durante o estágio de treinamento.

Existem outras possibilidades para definir **imersões relativas** semelhantes de itens de um espaço de características heterogêneo em um espaço de distância relativa unidimensional ponto-a-classe, usando DTW a fim de preservar a coerência temporal das medidas relativas. Uma dessas alternativas é o uso da distância H de Hausdorff, definida para conjuntos gerais A, B e qualquer distância dada d como:

$$H(A, B) = \max \left(\sup_{x \in A} \inf_{y \in B} d(x, y), \inf_{x \in A} \sup_{y \in B} d(x, y) \right), \quad (3.7)$$

isto é, a distância entre os conjuntos é a maior distância que se é forçado a percorrer de algum ponto de um conjunto até o ponto mais próximo do outro conjunto. Quando um dos conjuntos é unitário, a expressão acima é simplificada para $H(x, B) = \sup_{y \in B} d(x, y)$ a partir do qual definimos uma imersão DTW relativa como

$$\text{HausdorffDTW}(x) = \max_{y \in \mathcal{N} \setminus \{x\}} \text{DTW}(x, y). \quad (3.8)$$

A principal motivação para considerar as distâncias de Hausdorff neste contexto de classificação é o fato de que os itens fora da classe são comparados com o item mais distante da classe de interesse possível, o que pode facilitar a tarefa de classificação. Também é verdade, no entanto, que as distâncias intra-classe aumentarão em geral, mas o quanto elas aumentarão depende da distribuição dos valores $\text{DTW}(x, y)$ para $x, y \in \mathcal{N}$. Por exemplo, se todos os valores acima estiverem muito próximos uns dos outros (não necessariamente próximos de zero), como ocorre tipicamente para versões de um mesmo sinal com deformações temporais, as imersões HausdorffDTW relativas dos itens intra-classe poderiam permanecer mais ou menos na mesma região. Deve-se notar que, como o DTW não satisfaz a desigualdade triangular, não é necessariamente verdade que quando $\text{DTW}(x, y)$

---

³Considerando o complemento de $\mathcal{N}$ como uma segunda classe.

é pequeno para quaisquer $x, y \in \mathcal{N}$ então os valores de $\text{HausdorffDTW}(x, w)$ e $\text{MinDTW}(x, w)$ estariam próximos para $w \notin \mathcal{N}$.

3.1.2 Classificador baseado em perfis de dissimilaridade realinhados

Nessa seção apresentamos o método de classificação baseado em perfis de dissimilaridade realinhados, *Time-Aligned Dissimilarity Profile Classifier (TADPC)*. Este classificador tem como objetivo combinar a informação de todas as diferentes curvas de dissimilaridades construídas a partir do áudio a ser classificado e dos áudios sem alteração de fala da base de treinamento.

As curvas de dissimilaridades são construídas a partir dos MFCCgramas dos áudios alinhados pelo algoritmo de DTW, permitindo classificar áudios com diferentes durações mantendo a sincronia entre os fonemas das diversas gravações. Todas estas curvas são então combinadas em um único perfil de dissimilaridade, a partir do qual extraímos uma métrica usada para a classificação.

A Figura 3.3 representa a construção da curva de dissimilaridade de um áudio em relação aos áudios da base de treinamento. Podemos separar a imagem em duas etapas (da esquerda para direita): construção dos perfis individuais de dissimilaridade a partir das matrizes de dissimilaridade usadas no DTW, e obtenção do perfil de dissimilaridade unificado usado pelo classificador.

A primeira etapa é realizada percorrendo-se os caminhos ótimos obtidos pelo algoritmo de Viterbi dentro das matrizes de dissimilaridade do DTW, representando os valores de dissimilaridade em função do índice do frame do áudio de teste. Essa indexação permite representar todas as curvas de dissimilaridade (de um mesmo áudio de teste em relação a todos os áudios sem alteração de fala de treinamento) em um mesmo domínio temporal, o que corresponde efetivamente a realinhar todos os áudios sem alteração de fala de treinamento em relação ao áudio de teste.

Figura 3.3: Etapas do classificador baseado em perfis de dissimilaridade realinhados

Para formar a curva final de dissimilaridade (segunda etapa) utilizamos um percentil para definir o valor de dissimilaridade de cada frame, considerando a distribuição dos valores de dissimilaridade em todas as curvas naquele frame. A escolha deste percentil se dá através da otimização do valor de F-measure usando os dados de treinamento.

Utilizando o perfil de dissimilaridade $(x_1, \dots, x_n)$ resultante da segunda etapa, definimos a métrica de classificação pela área abaixo da curva

$$TADPC(x) = \frac{1}{N} \sum_{i=1}^N x_i \quad (3.9)$$

que será utilizada para classificar os áudios de teste, com ou sem alteração de fala, de acordo com um limiar também definido por otimização durante a etapa de treinamento.

Na Figura 3.4 podemos notar a diferença das curvas de um áudio sem alteração de fala com um áudio com alteração de fala. Embora ambos apresentem padrões variáveis, os picos do áudio sem alteração de fala alcançam valores de dissimilaridade muito menores do que o áudio com alteração de fala, resultando em valores de área também menores. A medida TADPC está representada pelas linhas vermelhas, que definem uma região retangular com área igual à da curva de dissimilaridade.

Figura 3.4: Comparação da curva de um áudio sem alteração de fala e de um áudio com alteração de fala

A definição do limiar de classificação a partir da medida TADPC segue o mesmo padrão dos demais classificadores apresentados. Começamos definindo as bordas das classes sem alteração de fala (N) e com alteração de fala (TF), através das expressões

$$\varrho^+ = \max_{x \in N} TADPC(x) \quad (3.10)$$

e

$$\varrho^- = \min_{y \in TF} TADPC(y) \quad (3.11)$$

Se as classes não possuem intersecção ($\varrho^+ < \varrho^-$), podemos definir o limiar como a média das duas bordas:

$$\tau = \frac{\varrho^+ + \varrho^-}{2}. \quad (3.12)$$

Caso contrário, podemos calcular o limiar ótimo de forma que ele maximize a medida de F-measure entre os dados de treinamento:

$$\tau = \underset{\alpha \in [\varrho^-, \varrho^+]}{\operatorname{argmax}} \text{ F-measure}(\alpha). \quad (3.13)$$

Utilizando este limiar, podemos classificar os áudios de teste de acordo com:

$$\begin{cases} z \in N & \text{se } TADPC(z) < \tau \\ z \in TF & \text{caso contrário.} \end{cases} \quad (3.14)$$

3.1.3 Vector Quantization + Bag-of-Words

O método denominado Vector Quantization + Bag-of-Words (VQ+BoW) começa com a etapa de clusterização (por exemplo usando K-médias) de todas as características conhecidas na etapa de treinamento, utilizando K clusters, e então cada item é codificado em um histograma K -dimensional, onde cada valor do histograma para $k = 1, 2, \dots, K$ representa quantos itens do vetor de características pertencem ao k -ésimo cluster. Desta forma, todos os itens de comprimento variável são representados através de características K -dimensionais homogêneas (e normalizadas).

Para a clusterização dos dados de treinamento, utilizamos a biblioteca “sklearn.cluster.KMeans”. Com isso é possível quantizar os vetores de características V com base no número de clusters, que é um parâmetro da função kmeans.

3.1.4 Modelo Oculto de Markov

Diferentemente do VQ+BoW, o HMM constrói um tipo diferente de representação, associando áudios sem alteração de fala a observações produzidas com alta probabilidade por um modelo

de Markov, onde K estados internos interconectados refletem a evolução temporal/estocástica do áudio, e as probabilidades de emissão de vetores de características fornecem o elo estatístico entre estados internos e vetores de características observados; A classificação prossegue então pela reconstrução de Viterbi de caminhos ótimos através da cadeia de Markov, com probabilidades associadas que permitem a discriminação entre áudios sem alteração de fala e com alteração de fala a partir das probabilidades obtidas pelo Viterbi: quanto maior a probabilidade maior a chance do áudio pertencer a classe das gravações sem alteração de fala.

Para implementar o HMM foi utilizada a biblioteca `hmmlearn` [hmm18], que busca seguir o mesmo padrão das bibliotecas do Scikit Learn.

3.1.5 Classificador baseado na *Earth movers distance*

O método de classificação utilizando a *Earth movers distance* (Seção 2.5.2) utiliza distribuições de probabilidade relativas aos valores de dissimilaridade obtidos no alinhamento dos áudios utilizando DTW. A *Earth mover's distance* se refere ao “esforço” de transformar uma função de densidade de probabilidade (f.d.p.) em outra, no nosso caso entre a distribuição dos valores de dissimilaridade produzidos por um áudio de teste e a distribuição dos mesmos valores produzidos por áudios sem alteração de fala.

O primeiro passo antes de implementar o classificador EMD é gerar a f.d.p. dos áudios sem alteração de fala da base de treinamento; para isso utilizamos todas as curvas de dissimilaridade entre todos os pares de áudios sem alteração de fala da base de treinamento. O segundo passo é calcular os valores de EMD entre a distribuição de cada áudio sem alteração de fala em relação à distribuição obtida de todos os áudios sem alteração de fala; isso definirá um conjunto de valores de EMD associados à classe dos áudios sem alteração de fala. Da mesma forma calculamos a EMD entre a distribuição de cada áudio com alteração de fala na base de treinamento e a distribuição de todos os áudios sem alteração de fala, produzindo os valores de EMD associados à classe de áudios com alteração de fala.

A Figura 3.5 mostra um exemplo de três funções de densidade de probabilidade, a primeira referente à distribuição de todos os sem alteração de fala (verde), a segunda referente a uma gravação sem alteração de fala específica (azul) e a última referente a uma gravação com alteração de fala (laranja). Ela ilustra a motivação do método, de que a distribuição obtida de um áudio sem alteração de fala fique muito mais próxima da distribuição de todos os áudios sem alteração de fala, quando comparada com uma distribuição obtida de um áudio com alteração de fala.

Figura 3.5: funções de densidade de probabilidade referentes à distribuição de todos os áudios sem alteração de fala (verde), a uma gravação sem alteração de fala específica (azul) e a uma gravação com alteração de fala (laranja), utilizadas no classificador baseado na *Earth movers distance*

A partir de todas as medidas EMD obtidas de áudios com e sem alteração de fala na fase de treinamento, conseguimos calcular um limiar que separa essas duas classes. Podemos escolher o limiar τ dentro do intervalo $[\mu_l, \mu_h]$, onde μ_l é o valor mínimo da EMD dos áudios da classe com alteração de fala e μ_h é o valor máximo da EMD dos áudios da classe dos sem alteração de fala.

Caso não haja intersecção entre as classes, ou seja $\mu_l > \mu_h$, definiremos o limiar como $\tau = \frac{\mu_l + \mu_h}{2}$. Cada áudio de teste t será então classificado de acordo com sua $EMD(t)$ em relação à distribuição de todos os áudios sem alteração de fala, sendo classificado como sem alteração de fala se $EMD(t) \leq \tau$, e como alteração de fala caso contrário.

3.2 Metodologia Experimental

3.2.1 Metodologia experimental MinDTW

Nesta seção apresentamos o desenho de um experimento piloto usando a base UA-Speech descrita na Seção 2.1.2. Na fase de extração de características nós utilizamos os coeficientes Mel-Cepstrais (MFCC) para representar cada quadro do áudio; MFCCgramas são representações matriciais bastante utilizadas no processamento de fala e que estão associadas ao conteúdo fonético [DM80], sendo por isso adequadas ao problema que queremos resolver, pois os áudios que contêm disartria são essencialmente modificações do conteúdo fonético em relação às falas sem alteração de fala, que por sua vez são muito similares entre si em termos dessa característica do áudio [JNM16]. Cada vetor de MFCC com 12 coeficientes é obtido de um frame de áudio de 2048 amostras (os frames têm sobreposição de 75%) utilizando a biblioteca librosa [MQ14].

Cada gravação i da base de dados UA-Speech é representada pelo MFCCgram M^i . Utilizando toda a base de dados de treinamento, nós construímos uma matriz de similaridade S por meio da distância DTW entre todos os pares de MFCCgramas, ou seja,

$$S_{ij} = DTW(M^i, M^j). \quad (3.15)$$

Figura 3.6: Matriz de similaridade da palavra “Seven” na base de dados UA-Speech. $\mathcal{N} = \{0, \dots, 39\}$ corresponde à classe dos áudios rotulados como sem alteração de fala e o restante é a classe dos áudios que apresentam alguma alteração de fala.

A Figura 3.6 mostra a matriz de similaridade das gravações referentes à palavra “SEVEN”, onde é possível notar que a classe dos áudios sem alteração de fala $\mathcal{N}$ (primeiros 40 itens) formam um *cluster* bem compacto (o quadrado verde no canto superior esquerdo da Figura indica que os valores de distâncias entre os membros desta classe são bem pequenos) enquanto que os áudios rotulados como contendo alguma alteração de fala estão distantes entre si e distantes também dos áudios rotulados como sem alteração de fala.

De acordo com a estratégia de **imersão DTW relativa** definida na Seção 3.1.1, a classe dos áudios sem alteração de fala será utilizada para definir o limiar de classificação. Para todas as gravações, definimos as distâncias de ponto-a-classe

$$\text{MinDTW}(i) = \min_{j \in \mathcal{N} \setminus \{i\}} S_{ij} \quad (3.16)$$

e

$$\text{HausdorffDTW}(i) = \max_{j \in \mathcal{N} \setminus \{i\}} S_{ij}. \quad (3.17)$$

A classificação é feita utilizando uma das duas métricas acima, através de um limiar que é definido a partir das bordas de cada classe. O limiar é otimizado dentro do intervalo $[\mu_l, \mu_h]$ de forma que ele maximize a *F-measure*, onde μ_l é o valor mínimo da distância inter-classe e μ_h é o valor máximo de distancia intra-classe. Caso não haja intersecção entre as classes, ou seja $\mu_l > \mu_h$, é utilizado o limiar $\frac{\mu_l + \mu_h}{2}$.

3.2.2 Metodologia experimental TADPC

Nesta seção apresentamos o experimento que foi realizado utilizando as bases de dados descritas na Seção 2.1. Primeiro utilizamos os coeficientes Mel-Cepstrais (MFCC) para representar cada quadro do áudio; nesta fase utilizamos os mesmos parâmetros utilizados na metodologia experimental do classificador MinDTW (Seção 3.1.1). O segundo passo é gerar as curvas de dissimilaridade, através do alinhamento via DTW do áudio a ser classificado com cada áudio sem alteração de fala da base de treinamento, conforme a Figura 3.7.

Figura 3.7: Resultado do alinhamento de um áudio com todos os áudios sem alteração de fala da base de treinamento

O próximo passo é combinar as curvas obtidas em uma única curva, usando um percentil dos valores das curvas em cada frame. Por fim, calculamos a integral (área da curva) do TADPC, que representa a dissimilaridade acumulada utilizada como uma medida de pertinência à classe dos áudios sem alteração de fala. A definição do limiar para classificação a partir dos valores TADPC utilizamos a mesma estratégia do MinDTW.

O classificador é calibrado escolhendo tanto o percentil que define o TADPC quanto o limiar de classificação de forma que eles maximizem a F-measure. Especificamente, nós realizamos uma otimização nas duas variáveis, com o percentil no intervalo $[0.1, 0.9]$ e o limiar no intervalo $[\mu_l, \mu_h]$, onde μ_l é o valor mínimo de dissimilaridade acumulada dos áudios da classe com alteração de fala e μ_h é o valor máximo de dissimilaridade acumulada dos áudios da classe sem alteração de fala. Caso não haja intersecção entre as classes, ou seja $\mu_l > \mu_h$, é utilizado o limiar $\frac{\mu_l + \mu_h}{2}$.

3.2.3 Metodologia experimental Vector Quantization + Bag-of-Words

A primeira etapa do bag-of-words consiste em representar os dados através de histogramas de dimensão K escolhida a priori. Para isso, quantizamos todos os MFCCs de cada áudio, variando o número K de *clusters* entre 20 e 70. Em seguida, é gerado um histograma para cada áudio, representando a frequência relativa de cada *cluster* associado aos frames daquele áudio. Desta forma, temos representações homogêneas dos áudios, baseadas nos histogramas K-dimensionais, e com elas podemos buscar um limiar que separe da melhor forma possível as duas classes. Para isso, calculamos o centroide da classe dos áudios sem alteração de fala e mapeamos todos os áudios em

um espaço unidimensional usando a distância de cada áudio até o centroide da classe dos áudios sem alteração de fala.

Analogamente ao que é feito na classificação usando as imersões por distâncias DTW relativas, para calcular o limiar entre as classes calculamos qual é a maior distância dos áudios sem alteração de fala até o centroide dos áudios sem alteração de fala, e depois calculamos a menor distância dos áudios com alteração de fala até o centroide dos áudios sem alteração de fala. De posse dessas bordas de cada classe, o limiar que irá definir se um áudio de teste será classificado como sem alteração de fala ou com alteração de fala será obtido pela otimização da F-measure, ou seja, para cada possível limiar entre as bordas das classes será computado o F-measure dos áudios de treinamento e selecionado o melhor limiar.

Após percorrer todos os k-folds, ou seja, depois que cada uma das dobras foi utilizada como base de teste uma vez, a F-measure global é calculada utilizando a fórmula 2.28 descrita na Seção 2.6.3.

3.2.4 Metodologia experimental Modelo de Markov oculto

Analogamente ao VQ+BoW, a primeira etapa é separar os dados em conjuntos de treinamento e teste. Para calcular o HMM passamos como parâmetros apenas os áudios com rótulo sem alteração de fala da base de treinamento, ou seja, consideramos o HMM como um modelo da classe de áudios sem alteração de fala, na expectativa de que áudios com alteração de fala sejam associados a probabilidades baixas de aderência àquele modelo.

Variando o número de estados de 5 a 60, treinamos o algoritmo de HMM para gerar as matrizes de probabilidades iniciais, de transição e de emissão. A partir de então, é possível calcular a probabilidade, através do algoritmo de Viterbi, de um áudio qualquer (com ou sem alteração de fala) pertencer àquele modelo. A biblioteca `hmmlearn` calcula o logaritmo da probabilidade computada pelo algoritmo de Viterbi para evitar erros numéricos associados a probabilidades muito pequenas, e devolve o valor absoluto do logaritmo da probabilidade do áudio pertencer ao modelo gerado, para permitir a interpretação dos valores como se fossem uma espécie de distância à classe dos sem alteração de fala, portanto quanto menor o valor produzido por essa implementação do Viterbi, maior a chance do áudio pertencer à classe dos sem alteração de fala.

Dessa maneira, é possível representar os dados de treinamento num espaço unidimensional baseado no valor de Viterbi, e com isso definir as bordas de cada classe da mesma forma que fizemos com VQ+BoW e MinDTW, e com isso calcular o limiar que otimiza a F-measure dos áudios de treinamento, para depois aplicá-lo nos áudios de teste. Assim como nos dois métodos anteriores, a F-measure global é calculada utilizando a fórmula 2.28 descrita na Seção 2.6.3 a partir do resultado de todas as dobras na validação cruzada.

3.2.5 Classificador baseado na *Earth movers distance*

Analogamente ao TADPC (Seção 3.1.2), utilizamos os coeficientes Mel-Cepstrais (MFCC) para representar cada quadro do áudio e geramos as curvas de dissimilaridade dos áudios sem alteração de fala da base de treino entre si, utilizamos estas curvas para gerar a função de densidade de probabilidade dos áudios sem alteração de fala da base de treino.

Para cada áudio a ser classificado, geramos as curvas de dissimilaridade utilizando o DTW entre os áudios a ser classificado com cada áudio sem alteração de fala da base de treinamento. Juntamos todas as curvas geradas e calculamos a função de densidade de probabilidade do áudio a ser classificado.

O próximo passo é calcular a *Earth movers distance* definida na Seção 2.5.2 entre a f.d.p. do áudio a ser classificado com o a f.d.p. dos áudios sem alteração de fala da base de treino para encontrar a distância ("esforço") entre as duas funções de densidade de probabilidade, utilizamos esta distância como uma medida de pertinência à classe dos áudios sem alteração de fala. Para definir o limiar para classificação a partir da *Earth movers distance* utilizamos a mesma estratégia do MinDTW.

O classificador é calibrado escolhendo o limiar de classificação que maximize a F-measure, utilizando uma otimização dentro do intervalo $[\mu_l, \mu_h]$, onde μ_l é o valor mínimo de dissimilaridade acumulada dos áudios da classe com alteração de fala e μ_h é o valor máximo de dissimilaridade acumulada dos áudios da classe dos sem alteração de fala. Caso não haja intersecção entre as classes, ou seja $\mu_l > \mu_h$, é utilizado o limiar $\frac{\mu_l + \mu_h}{2}$.

Capítulo 4

Resultados experimentais

Este trabalho contou com o apoio do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP, que disponibilizou duas bases de dados. A primeira foi uma amostra com 20 áudios da palavra “chuva”, sendo 10 de pessoas sem alteração de fala e 10 de pessoas com distúrbio. A segunda foi a base de dados gravada durante o trabalho de mestrado da aluna Danira Francisco [FW17], composta de duas palavras (“chave” e “sapo”) em que 21 crianças participaram da coleta de áudio, totalizando aproximadamente 100 gravações de cada palavra, sendo que 60% foram rotuladas como sem alteração de fala e 40% rotuladas com alteração de fala na produção dos fonemas [s] e [ʃ]. Além destas, também foi utilizada a base pública UA-Speech [KHJP⁺08].

Para calcular a acurácia do modelo proposto, foi utilizada a F-measure com validação cruzada (Seção 2.6.3). Vamos separar este capítulo em duas seções, uma para os resultados obtidos utilizando a base de dados da FMUSP e outra para os resultados obtidos utilizando a base de dados UA-Speech. Todos os métodos foram treinados e testados seguindo os mesmos processos, de acordo com as mesmas diretrizes gerais:

- Cada palavra do banco de dados possui um certo número de gravações de referência (registros sem alterações de fala), sendo que cada coleção de gravações de referência e cada técnica de representação/classificação (DTW, VQ+BoW, HMM) define um modelo de classificação diferente.
- Todos os limiares de classificação são selecionados de forma a otimizar a performance sobre os dados de treinamento, independentemente para cada método.
- todos os métodos foram submetidos à validação cruzada com 10 folds, utilizando a biblioteca `sklearn.cluster.KMeans`. Para cada fold k , 90% dos dados foram usados para treinamento e 10% para teste, e as quantidades correspondentes de verdadeiros positivos TP[k], negativos verdadeiros TN[k], falsos positivos FP[k] e falsos negativos FN[k] foram armazenadas.
- A performance é calculada utilizando a fórmula global da F-measure (equação 2.28).

4.1 Resultados utilizando as bases de dados do Departamento de Fisioterapia, Fonoaudiologia e Terapia Ocupacional da FMUSP

A primeira base de dados fornecida pela FMUSP foi utilizada para validar a performance do modelo de imersão por distância DTW relativa (MinDTW) apresentado na seção 3.1.1, comparado com os modelos de VQ+BoW (seção 3.1.2) e HMM (seção 3.1.3).

As Figuras 4.1, 4.2, 4.3 e 4.4 ilustram o resultado de uma dobra da validação cruzada da classificação da palavra “Chuva” utilizando os métodos MinDTW, HausdorffDTW, VQ+BoW e HMM, respectivamente. O eixo horizontal representa a menor distância de cada item em relação às gravações sem alteração de fala usadas no treinamento (exceto o próprio item), o eixo vertical

representa os valores da função de densidade de probabilidade (f.d.p), a linha azul representa os dados sem alteração de fala e a linha vermelha os dados com alteração de fala. Os dados sem alteração de fala utilizados para treinamento são representados como círculos azuis e os áudios de treinamento com alteração de fala são representados como quadrados vermelhos. Os dados de teste são representados como triângulos, sendo que a cor azul representa dados sem alteração de fala e a cor vermelha os dados com alteração de fala.

Para os dados correspondentes às Figuras 4.1 a 4.4, os valores de F-measure são 100%, 66,6%, 100% e 66,6% respectivamente. Porém tais medidas de F-measure são específicas a uma única dobra e estão sujeitas a fortes vieses, decorrentes entre outros do fato da base de dados ser muito pequena, e de cada dobra contar apenas com duas amostras de teste. Assim, podemos ver que o desempenho de 100% do VQ+BoW na Figura 4.3 não é muito representativo, pois as duas classes aparecem praticamente sobrepostas, mas para sorte do método os dois dados de teste (os triângulos) aparecem cada um do lado “certo” do limiar. Apenas no caso do MinDTW (Figura 4.1) existe uma separação clara entre os exemplos sem alteração de fala (azuis) e com alteração de fala (vermelhos).

Figura 4.1: Classificação da palavra “Chuva” usando imersão MinDTW relativa. Círculos e quadrados representam os dados de treinamento e triângulos os dados de teste; o eixo horizontal é referente à distância relativa até a classe sem alteração de fala; linhas verticais representam (da esquerda para direita) ϱ^+ , τ e ϱ^- .

Figura 4.2: Classificação da palavra “Chuva” usando imersão HausdorffDTW relativa. As convenções utilizadas são as mesmas da Figura 4.1.

Na Figura 4.5 comparamos o resultado da F-measure global dos 4 métodos para esse mesmo experimento com a palavra “Chuva” do banco de 20 gravações da FMUSP. Nessa avaliação global, o método de MinDTW teve o melhor resultado obtendo 100% de acurácia, ou seja este método

Figura 4.3: Classificação da palavra “Chuva” usando VQ+BoW.

Figura 4.4: Classificação da palavra “Chuva” usando HMM.

conseguiu classificar corretamente as duas classes em todas as dobras. Os demais métodos obtiveram F-measures de 90% (VQ+BoW), 66.6% (HMM) e 61.5% (HausdorffDTW). Vale lembrar que, devido ao pequeno número de amostras neste experimento, os resultados dos métodos em bases maiores devem mudar.

Figura 4.5: Valor da F-measure global dos métodos MinDTW, HausdorffDTW, VQ+BoW e HMM

A segunda base foi utilizada para validar os modelos de imersão MinDTW relativa (seção 3.1.1), TADPC (seção 3.1.4) e o classificador baseado em EMD (seção 3.1.5).

A tabela 4.1 apresenta duas medidas de F-measure, F-measure global que foi definida na seção 2.6.3 e a média/desvio padrão da F-measure de cada dobra individual. É possível notar que o classificador baseado em EMD teve o pior desempenho. Uma possível explicação para este resultado é o fato de que o EMD, assim como o VQ+BOW, descarta a sequência temporal do

áudio; em outras palavras, ele compara apenas as funções de densidade de probabilidade independentemente da ordem em que as dissimilaridades aparecem na comparação entre dois áudios.

A fim de comparar os métodos remanescentes, foi aplicado o teste t de Student pareado nas F-measures que foram obtidas nas K dobras da validação cruzada, utilizando a função `ttest_rel` do *Scipy*. Baseado no resultado dos testes, as únicas diferenças estatisticamente significantes foram que TADPC e HausdorffDTW performaram melhor que o MinDTW para a palavra “chave” ($p \ll 0.01$).

	sapo (['sapu'])		chave (['favi'])	
	F-GLOBAL (%)	F-NORMAL(%)	F-GLOBAL (%)	F-NORMAL(%)
MinDTW	81.42	81.34 ± 0.07	75.18	75.16 ± 0.03
HausdorffDTW	80.7	80.62 ± 0.06	76.06	76.08 ± 0.02
EMD	76.78	76.33 ± 0.1	64.66	59.23 ± 0.25
TADPC	81.03	80.78 ± 0.05	76.26	76.13 ± 0.03

Tabela 4.1: F-measure global e média/desvio padrão das F-measures das dobras individuais, para cada método e palavra.

Figura 4.6: Todas as curvas de dissimilaridade criadas a partir de uma gravação com alteração de fala da palavra “chave” comparada com as gravações sem alteração de fala da base de treino.

Figura 4.7: Mesma análise da Figura 4.6 utilizando outra gravação com alteração de fala da palavra “chave”.

Uma possível explicação para a performance superior do TADPC em relação ao MinDTW para a palavra “chave” pode estar relacionada às alterações de fala encontradas nesta palavra,

que costumam ser mais evidentes no julgamento perceptual, apesar da existência de alguns poucos *outliers* que, mesmo contendo alterações de fala, possuem valores de dissimilaridade em geral mais baixos. Estes *outliers* afetam o MinDTW pois esse método sempre vai utilizar o menor valor de dissimilaridade acumulado em relação às gravações de referência, aumentando a chance destes áudios com alteração de fala serem classificados como sem alteração (falsos positivos). Veja por exemplo na Figura 4.6, onde a maioria das curvas possui valores altos de dissimilaridade entre os frames 5 e 18, mas as curvas destacadas nas cores verde e roxo mantêm-se na região de valores baixos. Nesses casos, o valor de dissimilaridade da gravação calculada pelo MinDTW seria referente à curva que representam o menor valor de dissimilaridade, o que neste caso seria um *outliers*, confundindo-o com uma gravação sem alteração, enquanto o TADPC considera os valores mais altos de dissimilaridade de acordo com o parâmetro de percentil α . Deve-se notar que um classificador baseado exclusivamente no pior perfil, o HausdorffDTW, também foi considerado nesse estudo, mas como veremos na próxima seção, sua performance foi bem pior do que o MinDTW na base de dados UA-Speech.

Figura 4.8: Todas as curvas de dissimilaridade criadas a partir de uma gravação com alteração de fala da palavra “sapo” comparada com as gravações sem alteração de fala da base de treino.

Figura 4.9: Mesma análise da Figura 4.8 utilizando outra gravação com alteração de fala da palavra “sapo”.

Por outro lado, quando olhamos as curvas de dissimilaridade associadas à palavra “sapo”, os valores com alta dissimilaridade tendem a se concentrar nos mesmos trechos das gravações, como ilustrado na Figura 4.8, e assim escolher uma única curva de dissimilaridade (como o MinDTW) ou utilizar o parâmetro de percentil para resumir as curvas de dissimilaridade (como o TADPC)

acaba fazendo pouca diferença no resultado da classificação e consequentemente no valor final da F-measure.

4.2 Resultados utilizando a base de dados UA-Speech

A base da UA-speech também foi utilizada em dois experimentos: o primeiro foi a comparação do classificador MinDTW apresentado na seção 3.1.1, comparado com os modelos de VQ+BoW (seção 3.1.3) e HMM (seção 3.1.4), e o segundo experimento foi a comparação dos resultados dos classificadores MinDTW e TADPC (seção 3.1.2).

As Figuras 4.10 e 4.11 mostram o resultado de uma dobra da validação cruzada da classificação das palavras “Command” (pertencente a classe de palavras de comando) e “Hypothesis” (pertencente a classe de palavras incomuns) utilizando o método de imersão DTW relativa e o mesmo código de cores e formas das figuras anteriores.

Figura 4.10: Classificação da palavra “Command” usando imersão DTW relativa.

É possível verificar na Figura 4.10 que mesmo os dados de treinamento não sendo separados perfeitamente, ou seja, que existe uma intersecção entre as duas classes, os dados de testes foram perfeitamente classificados a partir do limiar calculado, o que permitiu neste exemplo em particular a obtenção de um F-measure de 100%, uma medida aparentemente enviesada. Na Figura 4.11 (exemplo da palavra “Hypothesis”) é possível ver um dado de teste (rotulado como sem alteração de fala) foi classificado como contendo alteração de fala, desta forma o F-measure para esta dobra foi de 66%.

Figura 4.11: Classificação da palavra “Hypothesis” usando imersão DTW relativa.

Figura 4.12: Classificação da palavra “Command” utilizando VQ+BoW; o eixo horizontal é a distância até o centroide da classe de palavras sem alteração de fala.

Figura 4.13: Classificação da palavra “Hypothesis” utilizando VQ+BoW.

Nas Figuras 4.12 e 4.13 é possível verificar que a distribuição dos dados sem alteração de fala e com alteração de fala utilizando o método VQ+BoW é muito sobreposto, tornando difícil a tarefa de separar as classes. Para a palavra “Command” foi obtido um F-measure de 75% e para a palavra “Hypothesis” o valor da F-measure foi 66%, nas respectivas dobras ilustradas nas Figuras 4.12 e 4.13.

Figura 4.14: Classificação da palavra “Command” usando HMM; o eixo horizontal é o valor absoluto do logaritmo da probabilidade de Viterbi.

Figura 4.15: Classificação da palavra “Hypothesis” usando HMM.

As Figuras 4.14 e 4.15 mostram as mesmas palavras sendo classificadas utilizando HMM. Novamente temos as duas classes sobrepostas, com F-measure de 91% para a palavra “Command” e 66% para a palavra “Hypothesis” nas respectivas dobras.

A Figura 4.16 mostra a comparação dos resultados obtidos (média e variância da F-measure global) dos algoritmos estudados (DTW, VQ+BoW e HMM) para todas as categorias de classes na base de dados UA-Speech original (considerando áudios de apenas um microfone). É visível que o algoritmo de imersão DTW relativa é superior aos métodos convencionais para este banco de dados. No geral o HMM possui um desempenho melhor que o VQ+BoW, o que já era esperado, pois este tipo de classificação é muito sensível à sequência temporal dos frames, um aspecto importante dos modelos dependentes do tempo, como o HMM e o DTW, mas que é totalmente ignorado pela abordagem do Bag-of-Words.

Figura 4.16: Valor da F-measure global utilizando a base de dados UA-Speech original.

Pode-se notar na coluna MinDTW da Figura 4.16 que a classe de palavras incomuns tem pior performance e também a maior variação. Uma possível explicação se deve ao fato de que esta classe possui palavras que podem apresentar uma certa dificuldade tanto para as pessoas com disartria quanto para as pessoas sem qualquer alteração de fala, distorcendo as bordas entre as classes. Além

disso, esta classe possui menos amostras (3 vezes menos que as outras classes), dificultando tanto o treinamento quanto o teste para esta classe de palavras.

A Figura 4.17 é uma extensão dos resultados obtidos na tabela 4.16. Nesta segunda tabela estamos utilizando a base UA-Speech estendida, onde para cada gravação temos 8 microfones posicionados em lugares diferentes. Comparando as colunas MinDTW e HausdorffDTW em ambas as tabelas, é possível dizer que HausdorffDTW produz uma classificação inferior ao MinDTW. Sendo orientado a refletir uma espécie de pior caso na distância entre conjuntos, a distância de Hausdorff aumenta as medidas imersas não só das gravações rotuladas como alteração de fala, mas também das gravações rotuladas como sem alteração de fala, que no caso desses dados se refletiu em um número maior de erros de classificação. Esse resultado não é suficiente para desconsiderar o uso da imersão HausdorffDTW, cuja utilidade, no caso geral, depende de quão próximos uns dos outros estão os itens da classe sem alteração de fala, em comparação com as medidas dos itens com alteração de fala.

Figura 4.17: Valor da *F-measure* global para os dados da base UA-Speech estendida.

Comparando os resultados gerais entre as duas tabelas, seria possível supor que a disponibilidade de mais gravações da mesma palavra reforçaria o algoritmo de classificação e o tornaria mais robusto às variações de timbre (a principal diferença de característica entre as oito gravações com microfones posicionados em locais diferentes).

O segundo experimento que realizamos utilizando a base de dados da UA-Speech foi comparar os algoritmos minDTW e TADPC. A Figura 4.18 mostra a comparação dos resultados obtidos (média e variância da *F-measure* global) para todas as categorias de classes na base de dados UA-Speech original (considerando áudios de apenas um microfone). É possível notar que o valor da *F-measure* do classificador MinDTW é superior ao da TADPC para esta base de dados.

Embora o classificador TADPC tenha tido uma *F-measure* comparável à do minDTW no experimento sobre a base de dados da FMUSP (Tabela 4.1), este resultado tão baixo na base UA-Speech pode ser explicado por dois fatores detalhados a seguir.

O primeiro fator corresponde ao fato de que as alterações de fala encontradas nas duas bases são bem diferentes: na base da FMUSP temos alterações de fala que correspondem à troca de fonemas que ocorrem no desenvolvimento da fala, enquanto nesta base encontramos alterações de fala que acontecem por conta da disartria, conforme apresentamos na seção 1.1.1, ou seja mesmo a pessoa falando a palavra de forma correta do ponto de vista fonológico, existem outras alterações na produção da fala.

O segundo fator está relacionado à forma como os dois classificadores funcionam. O TADPC busca criar uma única curva de dissimilaridade a partir da combinação das curvas obtidas através

Figura 4.18: Valor da *F-measure* global para os dados da base *UA-Speech*.

do alinhamento entre o áudio a ser classificado e cada áudio sem alteração de fala da base de treinamento. Por outro lado o minDTW busca o menor valor de dissimilaridade acumulado entre o áudio a ser classificado e os áudios sem alterações de fala da base de treino. Desta forma, quando ocorre uma troca de fonema, o classificador TADPC vai observar um valor maior de dissimilaridade naqueles frames, aumentando o valor da área abaixo desta curva em relação a um áudio sem alteração de fala (Figura 3.4 apresentada na seção 3.1.2). Por outro lado, se existe um problema de velocidade na produção da fala, o valor de dissimilaridade vai ser menor, e ainda que se estenda por mais frames, não afetará a área abaixo da curva de dissimilaridade por conta da normalização da duração. Assim, o valor calculado pelo TADPC para um áudio com disartria pode não ser tão alto a ponto de permitir sua diferenciação em relação a um áudio sem alteração de fala (Figura 4.19).

4.3 Discussão

Uma das dificuldades encontradas nessa pesquisa é o fato que, em todas as bases de dados utilizadas, os rótulos de áudios com alteração de fala e áudios sem alteração de fala são dados de acordo com o diagnóstico da pessoa, e não da palavra específica que corresponde à gravação, ou seja, se uma pessoa tem alguma alteração de fala relacionada a um fonema que ocorre em uma palavra específica, mas ela consegue pronunciar uma outra palavra do banco de forma correta, ela ainda terá o rótulo daquela gravação marcado como contendo alteração de fala. Isso implica em uma situação em que um áudio sem alteração de fala seria usado como áudio com alteração de fala para efeito de treinamento no algoritmo de classificação, aumentando o número de Falsos Negativos (FN), pois a borda da classe de áudios com alteração de fala seria empurrada em direção à classe sem alteração de fala, diminuindo o $\text{Recall} = \frac{TP}{TP+FN}$ e consequentemente o F-measure.

Outra dificuldade encontrada nos testes é relacionada à falta de pré-processamento do áudio; uma parte significativa de ambas as bases de dados possui algum tipo de ruído, outras vozes na gravação ou trechos de silêncio no início ou no fim do áudio. O pré-processamento do áudio não é uma das prioridades do estudo, então as gravações que possuíam algum destes tipos de problema foram retiradas de forma manual a fim de evitar um aumento significativo no raio da

Figura 4.19: Valor da *F-measure* global para os dados da base *UA-Speech*.

classe sem alteração de fala e com isso aumentar o número de falsos positivos (FP), afetando a $\text{Precisão} = \frac{TP}{TP+FP}$ e consequentemente a *F-measure*. Idealmente, tais gravações deveriam ser tratadas automaticamente a fim de eliminar os aspectos prejudiciais à classificação, sem descartar a informação central.

Na base de dados da FMUSP coletada durante o trabalho de mestrado da aluna Danira Francisco [FW17] foi necessário realizar uma segmentação manual para remover ruídos. Especificamente, foi necessário remover uma parte inicial do áudio que correspondia ao som de um apito que era usado para instruir os pacientes a pronunciar a palavra do teste, que ocorria em algumas gravações (mas não todas).

Um aspecto digno de menção está relacionado a como algumas técnicas criam representações alternativas a partir dos dados originais, particularmente usando quantização das características MFCC, o que invariavelmente introduz ruído nos experimentos. O método proposto de imersão DTW relativa, que calcula as distâncias diretamente a partir dos MFCCgramas, não requer nenhum passo de quantização antes da classificação real. Bag-of-Words só faz sentido em um cenário quantizado, devido à própria natureza de como os histogramas são construídos; portanto, há uma relação intrínseca entre a escolha do parâmetro de quantificação K (usado no algoritmo K médias) e o desempenho da estratégia de classificação baseada em BoW. O HMM pode ser definido sobre os dados originais, usando modelos contínuos para as probabilidades de emissão; essa foi a abordagem adotada aqui. Alternativamente, os MFCCgramas poderiam ser quantizados para produzir modelos HMM discretos menores e possivelmente computacionalmente mais eficientes para as probabilidades de emissão. Experimentos preliminares sugeriram que as taxas de sucesso foram maiores usando os dados originais, então esse foi o padrão adotado em nosso experimento comparativo.

MinDTW e TADPC são métodos originais e foram propostos especificamente para explorar o caráter temporal ou sequencial das representações associadas às gravações de palavras pronunciadas em testes para detecção de alteração de fala fonológicos. Ambos dependem do alinhamento entre pares de áudio realizado pelo DTW, o que permite a consideração de medidas de dissimilaridade que são tomadas ao longo de um caminho específico, ou seja, medidas que só comparam trechos correspondentes nos dois áudios.

Considerando que VQ+BoW descarta a sequência temporal, temos que essencialmente qualquer reordenação de um mesmo MFCCgrama produziria os mesmos centroides e os mesmos histogra-

mas. Isso pode representar um enorme problema ou nenhum problema, de acordo com o contexto da aplicação: modelos para timbre global [APS05], reconhecimento de acorde [CWB10] ou identificar um falante [MQ14] são geralmente baseados em representações estatísticas que prescindem da sequência temporal exata, mas o reconhecimento de palavras [GY07], o reconhecimento de melodias [DBP⁺07] e a identificação de versões musicais (cover-songs) [SGH10] são tarefas dependentes do tempo, para as quais o reordenamento das matrizes de características destrói parte importante da informação a ser classificada.

O classificador baseado em *Earth Movers Distance* também compara funções de densidade de probabilidade, o que o aproxima ao VQ+BoW. Entretanto, os modelos de probabilidade construídos utilizam valores de dissimilaridade que ocorrem ao longo do caminho DTW que realinha dois áudios, ou seja, os modelos probabilísticos respeitam suas sequências temporais.

Os modelos HMM, analogamente ao MinDTW e TADPC, preservam implicitamente o nexo temporal dos MFCCgramas no treinamento das cadeias de Markov, uma vez que as probabilidades de transições nas cadeias refletem pares de MFCC adjacentes nas gravações. Entretanto, as diferenças de comprimento das gravações têm um impacto nas reconstruções de Viterbi, no sentido que se uma gravação passa muitos quadros em um único nó da cadeia de Markov, ela terá uma probabilidade reduzida em relação a um sinal similar que se move através da cadeia de Markov em um ritmo mais rápido, ou seja, um modelo HMM penaliza gravações mais longas em relação a outras foneticamente idênticas porém mais curtas.

Capítulo 5

Considerações finais

5.1 Conclusão

Este trabalho teve como objetivo analisar os modelos de classificação de fala mais utilizados e medir sua performance quando aplicados numa base de dados contendo gravações de pessoas com e sem alteração de fala. Para isso estudamos técnicas de classificação de fala baseadas em Coeficientes Mel-Cepstrais (MFCC) combinados com Cadeias de Markov escondidas (HMM), Bag-of-words (BoW) e *Earth Movers Distance*. Além disso, foram propostos dois novos algoritmos: o MinDTW e o TADPC.

Os classificadores foram testados em três bases de dados: duas bases contendo palavras em português fornecidas pela FMUSP e uma contendo palavras em inglês. Para medir a qualidade e comparar o resultado dos classificadores utilizamos a medida de *F-measure*. Dentre os cinco classificadores, podemos destacar dois deles: o MinDTW, que teve a maior *F-measure* quando aplicado na base de dados da *UA-Speech*, com valores acima de 95%; e o TADPC, que teve a melhor performance quando aplicado na base de dados da FMUSP, com valores acima de 76%.

Podemos afirmar o objetivo inicial do trabalho foi atingido, pois além de aplicar métodos conhecidos para o problema de classificação de alterações de fala, propusemos dois métodos originais que superaram os métodos tradicionais em duas bases de dados, uma em língua portuguesa e outra em língua inglesa.

O experimento comparando a performance do classificador MinDTW com Cadeias de Markov escondidas (HMM) e *Bag-of-words* (BoW) na base de dados da *UA-Speech* foi apresentado e publicado na Sound and Music Computing Conference (SMC 2018), que aconteceu em julho de 2018 em Limassol (Chipre). Além disso, o experimento realizado na base de dados da FMUSP comparando os classificadores MinDTW, TADPC e *Earth Movers Distance* foi aprovado e será publicado na Sound and Music Computing Conference (SMC 2020), que acontecerá em junho de 2020 em Turim (Itália).

5.2 Trabalhos futuros

Nesta seção apresentamos uma proposta de segmentador de áudio, cuja ideia é analisar apenas os áudios com alterações de fala a fim de encontrar o momento em que a alteração de fala acontece. Com o intuito de auxiliar no processo de triagem da fonoaudiologia e visto que conseguimos construir um classificador que tem uma medida de *F-measure* alta, o próximo passo é localizar o momento em que ocorre a alteração na fala do paciente, com isso poderemos diminuir ainda mais o tempo gasto no processo de triagem., desta forma o fonoaudiólogo conseguiria analisar apenas o trecho com alteração de fala ao invés de ter que ouvir o áudio completo.

A metodologia do segmentador é a mesma do classificador TADPC (seção 3.1.2) mudando apenas o último, ao invés de utilizar a área abaixo da curva de dissimilaridade como medida de pertinência a classe sem alteração de fala, este método busca encontrar um trecho do áudio cuja área seja superior à área total do áudio inteiro. (TADPS - Time-aligned Dissimilarity Segmenter)

Para evitar que o método encontre um segmento de áudio com tamanho muito pequeno que representa um trecho tão curto do áudio que é inaudível, por exemplo escolher o segmento de tamanho 1 que possui o valor máximo de dissimilaridade do áudio. Multiplicamos o valor da área do segmento por uma função constante, neste caso a raiz cúbica do número de quadros que o segmento possui. A raiz cúbica foi escolhida para que segmentos de áudio pequenos sejam "punidos" enquanto que para segmentos de áudios grandes sejam multiplicados por uma constante de valor próximo. Por fim o método escolhe o trecho do áudio que possui a maior área.

$$TADPS(x) = \sqrt[3]{n} \times \frac{1}{n} \sum_{i=1}^n x_i \quad (5.1)$$

A Figura 5.1 mostra uma matriz com todos os valores de áreas encontrados para todos pares i, j onde $i < j$ e $j < N$, quanto mais clara a região maior o valor da área, desta podemos ver os segmentos que podem representar o momento em que ocorre a alteração.

Figura 5.1: Matriz com os valores área dos segmentos dos áudios

A Figura 5.2 mostra o resultado do segmentador TADPS, destacando entre as linhas pretas o segmento do áudio que pode possuir alteração de fala. Por fim, a Figura 5.3 mostra uma análise feita pelo praat [BW19] que mostra o espectrograma do áudio completo (esquerda) e o áudio segmentado (direita), neste exemplo temos um paciente dizendo a palavra chave, porém com alteração no primeiro fonema.

Figura 5.2: Resultado da segmentação utilizando TADPS.

Figura 5.3: *Análise do áudio utilizando o software praat, a esquerda temos a representação do áudio inteiro e a direita do áudio segmentado.*

Referências Bibliográficas

- [APS05] J-J Aucouturier, François Pachet e Mark Sandler. The way it sounds: timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005. 40
- [ASH18] ASHA. Speech disorder dysarthria. <https://www.asha.org/public/speech/disorders/dysarthria/>, 2018. Último acesso em 17/04/2018. 2
- [BBL02] Peter L Bartlett, Stéphane Boucheron e Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002. 16
- [Bul07] J. Bullock. Libxtract: A lightweight library for audio feature extraction. 01 2007. 5
- [BW19] P. Boersma e D. Weenink. Praat: doing phonetics by computer [computer program]. <http://www.praat.org/>, 2019. Último acesso em 05/12/2019. 42
- [CAHY09a] L. S. Chee, O. C. Ai, M. Hariharan e S. Yaacob. Automatic detection of prolongations and repetitions using lpcc. Em *2009 International Conference for Technical Postgraduates (TECHPOS)*, páginas 1–4, Dec 2009. 6
- [CAHY09b] L. S. Chee, O. C. Ai, M. Hariharan e S. Yaacob. Mfcc based recognition of repetitions and prolongations in stuttered speech using k-nn and lda. Em *2009 IEEE Student Conference on Research and Development (SCORED)*, páginas 146–149, Nov 2009. 6, 12
- [CAMYSC12] Ooi Chia Ai, Hariharan M, Sazali Yaacob e Lim Sin Chee. Classification of speech dysfluencies with mfcc and lpcc features. 39:2157–2165, 02 2012. 6
- [CCF⁺12] Heidi Christensen, Stuart P. Cunningham, Charles Fox, Phil D. Green e Thomas Hain. A comparative study of adaptive, automatic recognition of disordered speech. Em *INTERSPEECH*, 2012. 6
- [CWB10] Taemin Cho, Ron J Weiss e Juan Pablo Bello. Exploring common variations in state of the art chord recognition systems. Em *Proceedings of the Sound and Music Computing Conference (SMC)*, páginas 1–8. Citeseer, 2010. 40
- [DBP⁺07] Roger B Dannenberg, William P Birmingham, Bryan Pardo, Ning Hu, Colin Meek e George Tzanetakis. A comparative evaluation of search techniques for query-by-humming using the musart testbed. *Journal of the Association for Information Science and Technology*, 58(5):687–701, 2007. 40
- [DM80] S. Davis e P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Em *IEEE Transactions on Acoustics, Speech and Signal Processing, (ASSP)*, páginas 357–366, Volume:28, 1980. 12, 25
- [DW79] L. Devroye e T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, September 1979. 16

- [EE96] P. Enderby e J. Emerson. Speech and language therapy: Does it work? *BMJ: British Medical Journal*, página 1655–1658, 10 1996. 1
- [Efr83] Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983. 16
- [Fri86] R. Goldman; M. Fristoe. *Goldman-Fristoe test of articulation*. American Guidance Service, 1986. 2
- [FS10] G. Forman e M. Scholz. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. Em *SIGKDD Explor. Newsl.*, páginas 49–57, Volume 12, Issue 1, 2010. vii, 18
- [FW17] Danira Francisco e Haydee Wertzner. Differences between the production of [s] and [ʃ] in the speech of adults, typically developing children, and children with speech sound disorders: An ultrasound study. *Clinical linguistics & phonetics*, 31:1–16, 01 2017. 9, 29, 39
- [Gei75] Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975. 16
- [GHP04] Aravind Ganapathiraju, Jonathan E Hamaker e Joseph Picone. Applications of support vector machines to speech recognition. *IEEE transactions on signal processing*, 52(8):2348–2355, 2004. 6
- [GPAKR00] Y.V Geetha, Karanth Pratibha, Rao Ashok e Shetty K Ravindra. Classification of childhood disfluencies using neural networks. 25:99–117, 06 2000. 6
- [GY07] Mark Gales e Steve Young. The application of hidden markov models in speech recognition. *Signal Processing*, 1(3):195–304, 2007. 40
- [HCAY12] M. Hariharan, Lim Sin Chee, Ooi Chia Ai e Sazali Yaacob. Classification of speech dysfluencies using lpc based parameterization techniques. *J. Med. Syst.*, 36(3):1821–1830, Junho 2012. 6
- [Her90] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990. 12
- [hmm18] hmmlearn. hmm learn user guide. <http://hmmlearn.readthedocs.io/en/latest/index.html#user-guide>, 2018. Último acesso em 08/06/2018. 11, 24
- [HP09] Bartrip J. Howell P, Davis S. The uclass archive of stuttered speech. journal of speech, language, and hearing research. Em *JSLHR*, páginas 556–569, 2009. 7
- [JNM16] G. Jhavar, P. Nagraj e P. Mahalakshmi. Speech disorder recognition using mfcc. Em *2016 International Conference on Communication and Signal Processing (ICCSP)*, páginas 0246–0250, April 2016. 1, 25
- [KHJP⁺08] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin e Simone Frame. Dysarthric speech database for universal access research. Em *Ninth Annual Conference of the International Speech Communication Association*, 2008. vii, 7, 9, 11, 29
- [KMRN08] R. Rajagopal K. M. Ravikumar, Balakrishna Reddy e H. C. Nagaraj. Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. Em *International Journal of Electrical and Computer Engineering*, volume 2, páginas 2142–2145, 2008. 6

- [Lar31] Selmer C Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45, 1931. [16](#)
- [Li87] Ker-Chau Li. Asymptotic optimality for c_{π}/c_L , c_L/c_L , cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987. [16](#)
- [Mah12] P. Mahesha. Vector quantization and mfcc based classification of dysfluencies in stuttered speech. 2:01–06, 09 2012. [6](#)
- [MQ14] Anderson Fraiha Machado e Marcelo Queiroz. A flexible and modular crosslingual voice conversion system. Em *Joint ICMC / SMC*, páginas 1312–1319, 2014. [25](#), [40](#)
- [MRL⁺15] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg e O. Nietok. librosa: Audio and music signal analysis in pytho. Em *In Proceedings of the 14th python in science conference*, páginas 18–25, 2015. [11](#)
- [MT87] Frederick Mosteller e John W Tukey. Data analysis, including statistics. *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1965-1986*, 4:601, 1987. [16](#)
- [Mül15] Meinard Müller. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015. [vii](#), [14](#), [15](#)
- [NBMM06] AL Navas, C Berlim, M Muszkat e MC Miranda. Neurodesenvolvimento e linguagem. *Berlim C, Muszkat M, Miranda MC, organizadores. Neuropsicologia de desenvolvimento. São Paulo: Memnon*, páginas 93–105, 2006. [1](#)
- [NNH⁺00] Elmar Noeth, Heinrich Niemann, Tino Haderlein, Michael Decher, Uwe Eysholdt, Frank Rosanowski e Thomas Wittenberg. Automatic stuttering recognition using hidden markov models. Em *Sixth International Conference on Spoken Language Processing*, volume 4, páginas 65–68, 01 2000. [6](#)
- [PA12] Stephanie Pancoast e Murat Akbacak. Bag-of-audio-words approach for multimedia event classification. Em *Thirteenth Annual Conference of the International Speech Communication Association*, páginas 2105–2108, 2012. [13](#)
- [PCY⁺07] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng e Yu-Yuan Lin. A comparative study of different weighting schemes on knn-based emotion recognition in mandarin speech. Em De-Shuang Huang, Laurent Heutte e Marco Loog, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, páginas 997–1005, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. [6](#)
- [PdOM11] Letícia Pimenta Costa Spyer Prates e Vanessa de Oliveira Martins. Distúrbios da fala e da linguagem na infância. *Revista Médica de Minas Gerais*, 21(4 Supl 1):S54–S60, 2011. [1](#)
- [PJSD13] K Poullose Jacob, Sunny Sonia e Peter S David. Combined feature extraction techniques and naive bayes classifier for speech recognition. 2013. [6](#)
- [PVG⁺11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. [11](#)
- [Rab89] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989. [13](#)

- [RRN09] K.M. Ravikumar, R Rajagopal e H.C. Nagaraj. An approach for objective assessment of stuttered speech using mfcc features. 9:19–24, 01 2009. 1, 6
- [RTG00] Yossi Rubner, Carlo Tomasi e Leonidas Guibas. The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 11 2000. 15
- [Rud82] Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, páginas 65–78, 1982. 16
- [SC78] H. Sakoe e S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. Em *IEEE Transactions on Acoustics, Speech, and Signal Processing*, páginas 43 – 49 Volume: 26, Issue: 1, Feb 1978, 1978. 14
- [Sci19] SciPy. Scipy reference. available at. <https://docs.scipy.org/doc/scipy/reference/>, 2019. Último acesso em 08/12/2019. 11
- [SGH10] Joan Serra, Emilia Gómez e Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. Em *Advances in Music Information Retrieval*, páginas 307–332. Springer, 2010. 40
- [Sha93] Jun Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993. 16
- [Shr85] Kwiatkowski J. Shriberg, L. D. Continuous speech sampling for phonologic analysis of speech-delayed children. *Journal of Speech and Hearing Disorders*, 50(4):323–334, 1985. 2
- [SHR01] FLIPSEN JR. P. KARLSSON H. B. MCSWEENEY J. L. SHRIBERG, L. D. Acoustic phenotypes for speechgenetics studies: An acoustic marker for residual /z / distortions. *Clin. Ling. Phonetics*, 15(8):631–650, 2001. 2
- [SHR02] L. D. SHRIBERG. Classification and misclassification of child speech sound disorders. *ANNUAL CONVENTION OF THE AMERICAN SPEECHLANGUAGE-HEARING ASSOCIATION*, 17(7):507–528, 2002. 2
- [SK82] Lawrence D Shriberg e Joan Kwiatkowski. Phonological disorders i: A diagnostic classification system. *Journal of Speech and Hearing Disorders*, 47(3):226–241, 1982. 9
- [SLD03] P. KWIATKOWSKI J. MCSWEENEY J. L. SHRIBERG L. D, FLIPSEN JR. A diagnostic marker for speech delay associated with otitis media with effusion: the intelligibility-speech gap. *Clin. Ling. Phonetics*, 17(7):507–528, 2003. 2
- [Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. 16
- [Sto84] Charles J Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, páginas 1285–1297, 1984. 16
- [THLA⁺07] T. S. Tan, Helbin-Liboh, A. K. Ariff, C. M. Ting e S. H. Salleh. Application of malay speech technology in malay speech therapy assistance tools. Em *2007 International Conference on Intelligent and Advanced Systems*, páginas 330–334, Nov 2007. 6, 12
- [WER02] H.F. WERTZNER. *O Distúrbio Fonológico em Crianças Falantes do Português: Descrição e Medidas de Severidade*. Tese de Doutorado, Faculdade de Medicina da Universidade de São Paulo, Dezembro 2002. 2

- [Wer03] H. F. Wertzner. *Procedimentos de avaliação e tratamento do distúrbio fonológico*. Guanabara-koogan, 2003. [2](#)
- [WER06] PAPP A. C. C. S. GALEA D. E. S. WERTZNER, H. F. Provas de nomeação e imitação como instrumentos de diagnóstico do transtorno fonológico. *Pró-Fono Revista de Atualização Científica*, 18(3):303–312, 2006. [2](#)
- [WER09] PAGAN L. O. GURGUEIRA A. L. WERTZNER, H. F. Influência da otite média no transtorno fonológico: análise acústica da duração das fricativas do português brasileiro. *Revista CEFAC*, 11(1):11–18, 2009. [1](#)
- [WHF04] Befi-Lopes D. M. Fernandes F. D. M. Wertzner H. F., Andrade C.R.F. *ABFW: Teste de Linguagem Infantil nas Áreas de Fonologia, Vocabulário Fluência e Pragmática*. Pró-Fono, 2004. [2](#)
- [WKJSS07a] Marek Wisniewski, Wiesława Kuniszyk-Józkowiak, Elzbieta Smolka e Waldemar Suszynski. Automatic detection of disorders in a continuous speech with the hidden markov models approach. páginas 445–453, 10 2007. [6](#)
- [WKJSS07b] Marek Wisniewski, Wiesława Kuniszyk-Józkowiak, Elzbieta Smolka e Waldemar Suszynski. Automatic detection of prolonged fricative phonemes with the hidden markov models approach. 11, 01 2007. [6](#)
- [WN76] G White e R Neely. Speech recognition experiments with linear predication, band-pass filtering, and dynamic programming. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(2):183–188, 1976. [12](#)