Rafael Menêses Santos

Uma abordagem híbrida CNN-HMM para reconhecimento de fala tolerante a ruídos de ambiente

São Cristóvão - SE
2016

Rafael Menêses Santos

Uma abordagem híbrida CNN-HMM para reconhecimento de fala tolerante a ruídos de ambiente

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Leonardo Nogueira Matos
Coorientador: Hendrik Teixeira Macedo

São Cristóvão - SE
2016

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

S237u

Santos, Rafael Menêses
Uma abordagem híbrida CNN-HMM para reconhecimento de fala tolerante a ruídos de ambiente / Rafael Menêses Santos; orientador Leonardo Nogueira Matos. – São Cristóvão, 2016.
37 f.: il.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Sergipe, 2016.

1. Computação. 2. Redes neurais (Computação). 3. Reconhecimento automático de voz. 4. Markov, Processos de. I. Matos, Leonardo Nogueira. II. Título

CDU: 004.822

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÂO E PESQUISA
NÚCLEO DE PÓS-GRADUAÇÂO EM CIÊNCIA DA COMPUTAÇÃO

Relatório de defesa pública do(a) Senhor(a) **RAFAEL MENÊSES SANTOS** no Programa de Ciência da Computação (PROCC) da UFS.

Aos 30 dias do mês de maio de 2016, realizou-se a **Defesa de Mestrado** do trabalho intitulado "**Reconhecimento de fala em ambientes ruidosos usando Redes Neurais Convolucionais**" sob orientação do Prof. Dr. **Leonardo Nogueira Matos.**

Depois de declarada aberta a sessão, o Presidente da Banca passou inicialmente a palavra ao candidato para exposição e a seguir aos examinadores para as devidas arguições que se desenvolveram nos termos regimentais. Em seguida, a comissão julgadora proclamou o resultado:

Nome Banca Examinadora	Instituição	Assinatura
Leonardo Nogueira Matos	UFS	
Jugurta Rosa Montalvão Filho	UFS	
Tsang Ing Ren	UFPE	

Dessa maneira o Resultado Final é: X **APROVADO** ou Reprovado

Parecer da Banca Examinadora *

* Obs: Se o candidato for reprovado, o preenchimento do parecer é obrigatório.

São Cristóvão/SE

Assinatura do Orientador:

Assinatura do Aluno:

Resumo

Um dos maiores desafios no reconhecimento de fala atualmente é usá-lo no contexto diário, no qual distorções no sinal da fala e ruídos no ambiente estão presentes e reduzem a qualidade do reconhecimento. Nos últimos trinta anos, centenas de métodos para reconhecimento robusto ao ruído foram propostos, cada um com suas vantagens e desvantagens. Este trabalho propõe o uso de uma rede neural convolucional no papel de modelo acústico em sistemas de reconhecimento automático de fala, como uma alternativa ao métodos clássicos de reconhecimento baseado em modelos ocultos de Markov (HMM, do inglês, *Hidden Markov Models*) sem a aplicação de um método robusto ao ruído. Experimentos foram realizados com áudios modificados com ruídos aditivos e reais, e mostraram que o método proposto reduz o *Equal Error Rate* (EER) e aumenta a acurácia da classificação de comando de voz quando comparado a modelos tradicionais de classificação, evidenciando a robustez da abordagem apresentada.

Palavras-chave: Reconhecimento de fala, Redes Neurais Convolucionais, HMM.

Abstract

One of the biggest challenges in speech recognition today is its use on a daily basis, in which distortion and noise in the environment are present and hinder this task. In the last thirty years, hundreds of methods for noise-robust recognition were proposed, each with its own advantages and disadvantages. In this thesis, the use of Convolutional Neural Networks (CNN) as acoustic models in automatic speech recognition systems (ASR) is proposed as an alternative to the classical recognition methods based on Hidden Markov Models (HMM) without any noise-robust method applied. Experiments were performed with a audio set modified by additive and natural noises, and showed that the presented method reduces the Equal Error Rate (EER) and improves the accuracy of speech recognition in noisy environments when compared to traditional models of classification, indicating the robustness of the approach.

Keywords: Speech Recognition, Convolutional Neural Networks, HMM.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ASR	Automatic Speech Recognition
CNN	Convolutional Neural Networks
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
SVM	Support Vector Machines
DTW	Dynamic Time Warping
EAE	Evento Acústico Elementar
MFCC	Mel-Frequency Cepstrum Coefficients

Sumário

Capítulo 1

Introdução

1.1 Problema

Nos últimos anos, *smartphones* têm sido os dispositivos de comunicação mais usados no mundo. Segundo Smith (2015), em 2015, aproximadamente 64% da população norte americana possuía um *smartphone*, um aumento de quase 100% quando comparado os 35% de usuários no início de 2011. Sua crescente capacidade computacional combinada com sua alta abrangência fizeram com que as pessoas permanecessem completamente conectadas e constantemente disponíveis. A interface entre o usuário e o *smartphone* é realizada por meio de toques na tela, o que muitas vezes se torna pouco intuitivo e natural. Por se tratar de uma forma mais natural de interação, muitas empresas começaram a permitir de forma mais limitada até o momento, o uso da interface de voz com o *smartphone*, como meio mais natural de fornecer comandos. Este é mais um novo desafio para a área de Reconhecimento Automático de Fala (*Automatic Speech Recognition*, ASR) devido às diversas dificuldades referentes ao uso da voz, como: ruídos no ambiente, tamanho e forma do trato vocal, respiração, etc. A técnica mais comum usada para a modelagem acústica em sistemas ASR é uma combinação de modelos ocultos de Markov (Hidden Markov Models, HMM), responsáveis por modelar a estrutura sequencial do sinal de voz, e modelos de misturas Gaussianas (Gaussian Mixture Models, GMM) para modelar a representação acústica de características extraídas a partir do sinal (Abdel-Hamid et al., 2012). Esta abordagem é facilmente afetada por variações da fala em conversas diárias.

Para realizar tarefas de reconhecimento de fala, extratores de características são usados com o objetivo de representar os dados de entrada em uma forma mais conveniente. Nos últimos 50 anos, pesquisadores nessa área vêm desenvolvendo várias formas de representação com bases em características anatômicas, psicológicas e acústicas da fala humana, para serem usadas como características no reconhecimento de fala. Entre elas, as mais usadas são: banco de filtros (Fbank), Coeficientes Mel-Cepstrais (*Mel-frequency Cepstrum Coefficients*, MFCC) (Davis and Mermelstein, 1980) e Coeficientes PLP (*Perceptual Linear Prediction*) (Hermansky, 1990). Essas características possuem ao menos dois fatores que os tornam bastante adequados na representação da fala: a

redução da dimensionalidade do sinal e a preservação de uma quantidade de informação suficiente para tarefas de classificação, com pouca perda no desempenho. Contudo, elas não apresentam robustez ao ruído e dependem da avaliação e interpretação do responsável por parametrizar e escolher o tipo de extrator de característica. Uma forma de abordar este tipo de problema envolve o uso de uma das centenas de técnicas que foram estabelecidas para o tratamento de ruído (Li et al., 2014). Essa decisão requer conhecimento dos prós e contras que cada método pode oferecer.

Ultimamente, redes neurais profundas (*Deep Neural Networks*) têm sido propostas como substitutas do HMM na modelagem acústica (Hinton et al., 2012). Entre as redes neurais profundas mais conhecidas, as redes neurais convolucionais (*Convolutional Neural Network*, CNN) têm provado que podem ser treinadas de forma mais rápida, além de conseguirem excelentes resultados, entre eles, a menor taxa de erro na base MNIST até o momento (Ciresan et al., 2011, 2012). As CNNs são usadas principalmente para tarefas na área de visão computacional (LeCun et al., 2004) e em problemas de série temporal (Lee et al., 2009), nos quais são feitas filtragens baseadas em convoluções através do eixo temporal, realizando uma extração implícita de características dos dados de entrada na forma natural (Abdel-Hamid et al., 2012). Como mostra o trabalho de LeCun et al. (1998), as CNNs são capazes de extrair implicitamente características dos dados na sua forma natural, lidando de maneira robusta com a ruídos e variações presentes nos dados.

Neste trabalho, é levantada a hipótese, detalhada na Seção 3.2, de que o uso de redes neurais convolucionais como modelo acústico em uma abordagem híbrida com o HMM, é uma solução robusta para superar problemas causados por ruídos de ambiente e variações produzidas pela voz. Estas variações ocorrem devido ao uso do reconhecimento de fala em ambientes não controlados, ou seja, a utilização da interface de voz no contexto do dia a dia. O foco deste estudo foi direcionado ao reconhecimento de palavras isoladas do português brasileiro, levando em consideração a influência de diferentes tipos de ruídos no desempenho da solução proposta. Desta forma, o trabalho apresenta uma abordagem ao reconhecimento de fala que combina CNN e HMM, baseada no trabalho de Abdel-Hamid et al. (2012), e verifica a robustez da mesma em áudios modificados com ruídos aditivos e naturais.

1.2 Trabalhos Relacionados

Nos últimos cinco anos, o reconhecimento de fala com CNN tem sido o foco de vários estudos. Contudo, foram encontrados poucos estudos que tratam diretamente da influência do ruído no desempenho da classificação com a CNN.

Uma versão modificada do CNN chamada de Gabor Convolutional Neural Network (GCNN) é apresentada no trabalho de Chang and Morgan (2014). A rede incorpora funções Gabor nos filtros do kernel de convolução. A partir do treinamento da rede usando versões limpas e ruidosas da base Wall Street Journal, o estudo apresentou uma melhoria de desempenho quando comparado a outras técnicas robustas ao ruído.

O uso de características robustas com a CNN e DBN ajudaram a melhorar as taxas

de reconhecimento na base Aurora4 no estudo de Mitra et al. (2014). Isso é reforçado por Huang et al. (2015) que mostra ainda a CNN como modelo mais apropriado que a DBN para o reconhecimento de fala.

Por último, CNN é aplicada no contexto de detecção de frases e comparada ao GMM e MLP. No seu estudo, (Soltau et al., 2013) utiliza uma base de áudio limpo que é distorcida pela transmissão em oito canais diferentes de rádio. A partir disto, ele realiza o treinamento usando tanto dados limpos, como modificados. CNN conseguiu melhores resultados em canais ruidosos, mas não melhorou as taxas de desempenho conhecidas para esta base em situações mais controladas, ou seja, com ruido menos intenso.

1.3 Objetivos

O objetivo geral deste trabalho é definir e avaliar o modelo híbrido entre o CNN e HMM para reconhecimento de fala em amostras de áudio modificadas com ruído aditivo e ruídos naturais da própria gravação. Para alcançar o objetivo geral, foram identificados os seguintes objetivos específicos:

- Analisar os métodos mais usados para realizar o reconhecimento automático de fala;
- Apresentar e avaliar uma abordagem para reconhecimento de fala, combinando redes neurais profundas, do tipo convolucional, e modelos ocultos de Markov para modelagem dinâmica do sinal de fala;
- Realizar experimentos em dados de áudio adquiridos através de dispositivos móveis, com presença de ruído de ambiente e com ruídos simulados após a aquisição, visando avaliar o desempenho da abordagem em situações de uso real.

1.4 Organização da Dissertação

A dissertação foi dividida em cinco capítulos. O primeiro capítulo é a introdução do trabalho e os demais capítulos são descritos como se segue:

- Capítulo 2: Apresenta conceitos fundamentais que envolvem o trabalho realizado. A princípio, uma introdução ao reconhecimento automático de fala é feita, com o intuito de fornecer um bom esclarecimento dos principais componentes de uma arquitetura ASR generalizada. Em seguida são apresentados conceitos das Redes Neurais Convolucionais, que são o foco da dissertação.
- Capítulo 3: Discute a idéia por trás de modelo híbrido entre HMM e RNA e como a CNN fará o papel de uma MLP tradicional neste modelo. O capítulo encerra com uma discussão da hipótese levantada neste trabalho, discutindo características fundamentais da CNN que podem melhorar o desempenho do problema discutido no trabalho.

- Capítulo 4: Detalha os experimentos que foram realizados, bem como as bases e métricas usadas. Além disso, apresenta resultados dos experimentos e uma discussão.
- Capítulo 5: Apresenta projetos que utilizam o modelo desenvolvido neste trabalho.
- Capítulo 6: O último capítulo conclui o trabalho destacando principais contribuições, limitações de pesquisa e sugestões de novos caminhos a serem seguidos em possíveis trabalhos futuros.

Capítulo 2

Referencial Teórico

Neste capítulo são apresentados conceitos fundamentais utilizados no trabalho. A primeira seção fornece uma visão geral sobre a arquitetura de um sistema ASR, seguida pela seção que trata das redes neurais convolucionais.

2.1 Reconhecimento de Fala

No processo de reconhecimento de voz, o sinal de voz é primeiramente preprocessado no chamado *front-end* do sistema, que extrai as características que serão usadas. A maioria dos sistemas ASR usa uma abordagem baseada em *frames* do áudio, que convertem o sinal de entrada em sequências de *frames* de características com igual duração. Características como MFCCs e Fbanks são alguns exemplos de possíveis representações que podem ser usadas neste processo (Davis and Mermelstein, 1980). Na Figura 2.1, é apresentado um diagrama que mostra uma arquitetura genérica de um sistema ASR.

O decodificador é o módulo do sistema que processa o sinal acústico, transcrevendo-o em uma sentença escrita. É composto basicamente pelos modelos acústicos, modelos linguísticos e o dicionário.

O modelo acústico é um modelo estatístico baseado nas características computadas no *front-end*. Normalmente, esse modelo é usado para calcular a verossimilhança da geração de observações acústicas a nível de fonemas. Geralmente são usados Modelos Ocultos de Markov e Modelos de Misturas Gaussianas como componentes do modelo acústico (Rabiner, 1989).

O dicionário mapeia cada palavra de uma determinada língua à sua representação fonética, ou seja, uma sequência de fonemas. Uma palavra pode ter algumas alternativas de pronúncia, sendo dessa forma necessário que no dicionário existam múltiplas entradas para a palavra. O dicionário é usado como uma forma de restringir as possíveis sequências de fonemas, pois sem o mesmo, certamente não seria possível produzir uma decodificação confiável do sinal.

O modelo linguístico é o segundo modelo estatístico presente no decodificador, e possui o objetivo de modelar a fonte probabilidade das possíveis sequências em um

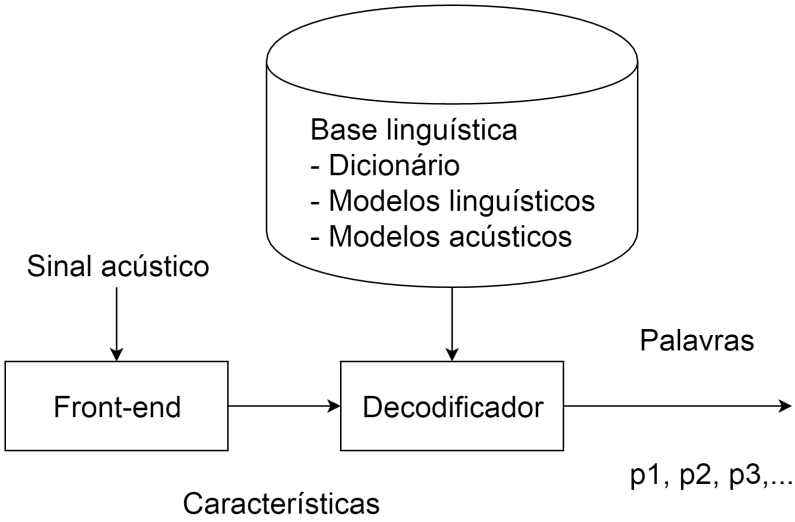

Figura 2.1: Arquitetura de um sistema ASR. Adpatado de (Rabiner, 1989)

idioma. Segundo Jurafsky and Martin (2000), a maioria dos sistemas ASR utiliza o modelo *N-gram*, devido à facilidade de combinação com outros componentes.

2.2 Redes Neurais Convolucionais

Com base no trabalho de Hubel and Wiesel (1968) sobre córtex visual dos gatos, podemos saber que essa estrutura contém um arranjo complexo de células. Essas células são sensitivas em pequenas regiões do campo de visão, chamadas de campos receptivos.

Dois tipos de células foram definidas: (i) As células simples, que são sensíveis a padrões específicos com forma de borda, e (ii) as células complexas, que possuem campos receptivos maiores e são localmente invariantes à posição exata do padrão.

O córtex visual animal é um dos sistemas de processamento mais poderosos conhecidos, e por isso, pesquisadores tentam simular seu funcionamento artificialmente. Alguns exemplos de modelos desenvolvidos são: NeoCognitron (Fukushima, 1980), HMAX (Serre et al., 2007) e Redes Neurais Convolucionais (LeCun et al., 1998).

Redes Neurais Convolucionais são variações de redes Perceptron Multicamadas (*Multilayer Perceptron Networks*, MLP) compostas por sucessivas camadas de convolução e subamostragem que realizam uma etapa de preprocessamento dos dados de entrada, e uma camada que corresponde a uma MLP, responsável por calcular a saída da CNN. A sua estrutura é modelada para tirar vantagem de problemas que envolvem padrões bidimensionais, tais quais imagens e sinais de fala em espectrograma. Espectrogramas são representações visuais do espectro de frequência em sinais em relação a sua variação no tempo.

A camada de convolução é composta por conjuntos de filtros, conhecidos como mapas de características, onde cada filtro atua localmente, simulando o comportamento

dos campos receptivos encontrados no sistema visual de organismos vivos (LeCun et al., 1998). Os filtros de cada mapa de características, também conhecidos como *kernels* de convolução, são entrelaçados e agrupados de tal forma que todo campo de entrada possa ser representado, além de considerar correlação entre neurônios vizinhos. Essa propriedade é conhecida como conectividade local e introduz robustez a deslocamento e distorções em amostras de mesma classe.

Outra propriedade importante da camada de convolução são os pesos compartilhados, definida como sendo o compartilhamento de pesos entre os *kernels* do mesmo mapa de características. Graças a esse compartilhamento de pesos, os mapas de características podem detectar padrões independentemente da localização no campo de entrada. Os pesos são estimados usando o treinamento com algoritmo de *backpropagation* modificado como sugere Abdel-Hamid et al. (2014), permitindo que cada mapa de característica possa encontrar um tipo particular de característica que é aprendida durante a fase de treinamento.

Após a camada de convolução, as ativações são passadas para uma segunda camada, a camada de subamostragem. Essa camada pode calcular o valor máximo ou médio de uma área de tamanho predefinido, gerando uma representação da entrada em resolução reduzida. Essa etapa melhora a invariância a translação, consequentemente, aumentando a acurácia da classificação LeCun et al. (1998). A arquitetura de uma CNN é apresentada na Figura 2.2.

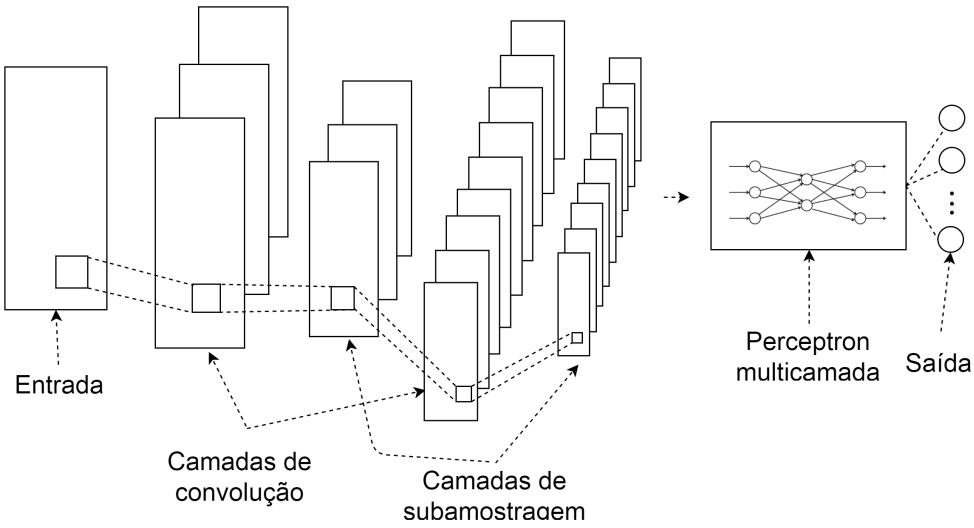

Figura 2.2: Arquitetura de uma CNN

Uma CNN pode conter uma ou mais camadas de convolução e subamostragem, conforme preferência e ajustes necessários relacionados ao problema de classificação. Como entrada, a CNN recebe uma matriz real $m \times n$ que é processada por uma camada de convolução. Em seguida, a camada de convolução é formada por k mapas de características com $c \times c$ *kernels*, onde $c < n$. Dado o *kernel* w, a saída y_{ij} do mapa de características m é calculada pela convolução da entrada x e w,

$$y_{i,j} = \sum_{a=-m}^{m} \sum_{b=-m}^{m} w_{a,b} x_{(i-a,j-b)}$$

onde o tamanho do *kernel* é igual a $(2m+1) \times (2m+1)$ (Abdel-Hamid et al., 2012). Após a convolução, uma função de ativação não linear é aplicada em cada mapa de características somada a um peso, definido como viés.

O próximo passo é processar a saída de cada mapa de características através de uma camada de subamostragem, que pode usar um filtro de média ou de valor máximo numa área predefinida de tamanho $r \times r$. Na Figura 2.3, uma representação das camadas de convolução e subamostragem é apresentada.

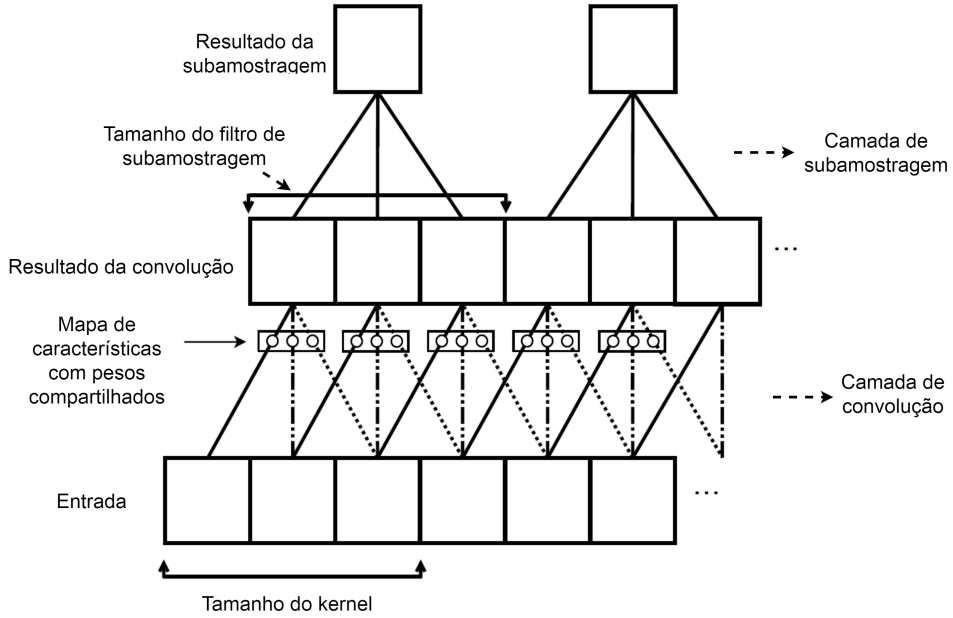

Figura 2.3: Camadas de convolução e subamostragem

Finalmente, o resultado das camadas iniciais é apresentado a uma MLP totalmente conectada que calcula a saída da CNN.

Capítulo 3

Modelo Proposto

3.1 Modelo CNN-HMM

Redes neurais podem ser usadas para classificar diversos padrões, entre eles, fonemas e palavras, mapeando um padrão temporal em espacial (Bourlard and Morgan, 1994). Entretanto, as redes neurais são limitadas em seu poder de classificar sequências, pois as mesmas podem variar infinitamente de tamanho. Por outro lado, o HMM possui uma estrutura capaz de lidar com sequências de tamanho indefinido.

Pensando nisso, um modelo híbrido NN-HMM foi criado, no qual a rede neural tem o papel de modelar uma representação acústica dos *frames* da fala, enquanto que o HMM modela a estrutura temporal e as dependências entre *frames* adjacentes. Cada entrada para rede é formada por uma sequência de T *frames* O_t, centralizados no tempo t. A saída da rede corresponde à probabilidade a *posteriori* $P(s|O_t)$ que representa o estado s do HMM no tempo t (Bourlard and Morgan, 1994). As probabilidades dos estados do HMM são calculadas na saída da rede pela função *softmax*

$$P(s|O_t) = \frac{exp(y_t)}{\sum_{i=1}^{T} exp(y_i)},$$

onde y_t é a saída do neurônio t. Desta forma, um HMM treinado encontra a sequência de estados s_1,s_2,...,s_n que melhor "explica" a sequência de observações O_1,O_2,...,O_n.

Neste trabalho, a rede neural MLP é substituída por uma rede neural convolucional. Esse modelo foi definido por Abdel-Hamid et al. (2012) e adpatado para o reconhecimento de fala conforme apresentado na Figura 3.1.

3.2 Detalhamento da Hipótese

A hipótese tratada neste trabalho está relacionada ao uso de Redes Neurais Convolucionais e como elas podem melhorar as taxas de acertos no reconhecimento de fala em ambientes ruidosos. O foco do estudo é direcionado ao reconhecimento de palavras isoladas na língua portuguesa, avaliando também a influência de diferentes tipos de ruídos

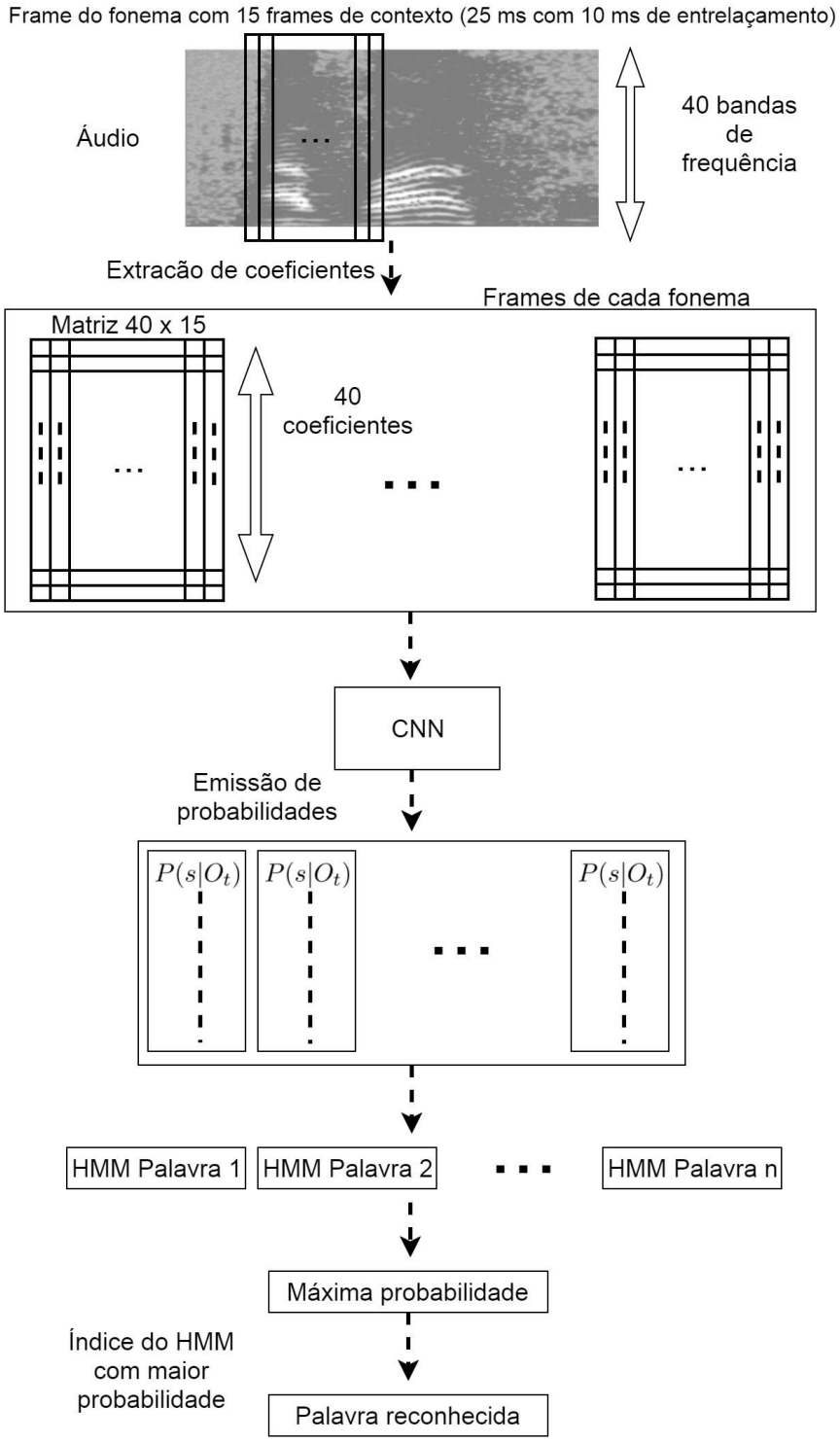

Figura 3.1: Modelo híbrido CNN-HMM (Abdel-Hamid et al., 2012)

no desempenho da solução proposta. A partir dos problemas mencionados na Seção 1.1, surge o questionamento: como superar os problemas causados pelos ruídos do ambiente e variação do sinal da fala, sem que haja dependência do processo de extração de características, utilizando diretamente os dados em sua forma natural?

No domínio da frequência, os sinais da fala são representado por concentrações de energias em diferente bandas de frequência. Quando um sinal de fala é analisado a nível fonético, é possível verificar que o fonema pode ser classificado a partir de sua representação espectral. Na presença de ruído, o sinal continua bastante representativo devido ao ruído estar concentrado apenas em algumas partes do espectro. De acordo com Abdel-Hamid et al. (2012), espectro na escala Mel, banco de filtros e espectro lineares são representações adequadas da entrada no domínio da frequência que preservam informações espaciais.

Uma vez que CNNs são especializadas em problemas que envolvem correlações espaciais, é esperado que os *frames* do sinal de fala no domínio da frequência possam ser usados como entrada para uma CNN, que por sua vez vai gerar uma sequência de observações ou probabilidades usadas pelo HMM. Em ambientes ruidosos, a CNN ainda pode extrair características representativas, porque os pesos compartilhados de cada mapa de características são replicados por todo o espaço da entrada, permitindo que características possam ser encontradas independente de sua localização, caso venham a ser distorcidas pelo ruído.

Em seu trabalho, LeCun et al. (1998) mostra como a CNN é capaz de superar variações e ruídos no reconhecimento de dígitos manuscritos. A Figura 3.2 inclui exemplos de robustez da CNN sob condições de ruídos diversos. Sem o uso de CNN, seria necessário submeter métodos de extração de características e segmentação da imagem.

No modelo apresentado, a CNN é treinada para estimar as probabilidades de emissões do HMM. Além de receber o *frame* que será classificado a nível fonético, a CNN recebe *frames* anteriores e posteriores, permitindo que ela possa aprender informações de contexto do fonema.

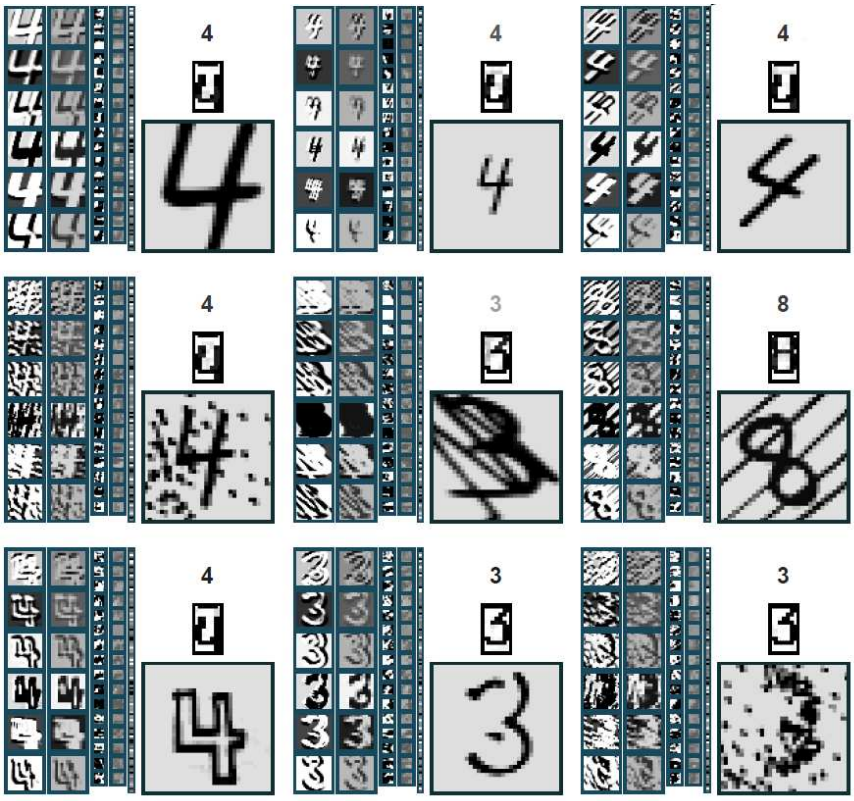

Figura 3.2: Exemplos de reconhecimento de imagem que apresentam robustez ao ruído. Fonte: LeCun et al. (1998)

Capítulo 4

Experimentos e Resultados

4.1 Cenários de Experimentação

Para avaliar o método proposto, foi conduzido um experimento que envolve o reconhecimento de palavras isoladas com ruído aditivo. O experimento é baseado no trabalho de Almeida (2014), no qual o reconhecimento de palavras dependente de locutor é avaliado em um banco de dados modificado com três ruídos da base NOISEX-92: conversa, volvo (denominação dada pelos autores ao ruído produzido pelo motor de um veículo específico) e fábrica (Varga and Steeneken, 1993).

Testes foram conduzidos para determinar a melhor configuração para o modelo. A configuração encontrada é composta por duas camadas de convolução e subamostragem. As camadas de convolução possuem *kernels* de tamanho 3×3 seguidas por camadas de subamostragem com filtros de tamanho 2×2. Na primeira camada, 20 mapas de características são usados e, na segunda, 50. A última camada é formada por uma MLP com 200 neurônios na camada oculta. Cada neurônio de saída representa um fonema que forma uma palavra a ser classificada.

O primeiro experimento trata basicamente de uma comparação mais direta entre CNN-HMM com modelos de classificação tradicionais (SVM e GMM). O SVM e GMM, nesse contexto, são aplicados na modelagem acústica dos sinais, assumindo o papel da CNN e utilizando o HMM para a modelagem sequencial. Além disso, é feita uma avaliação das características geradas pelo CNN em comparação com Eventos Acústicos Elementares, características propostas por Almeida (2014). Entretanto, este experimento difere bastante do que poderia ser encontrado em situações de uso cotidiano de um sistema ASR. Os principais motivos são:

- A utilização de treinamento e teste dependente de locutor: O experimento não verifica como a abordagem se comportaria num cenário mais diversificado, em que os locutores presentes na base de teste são completamente diferentes daqueles que participaram da construção da base de treinamento.
- Ruidos aditivos: Quando utilizamos ruídos aditivos, não consideramos algumas características que estão presentes em ambientes ruidosos naturais. Em ambiente

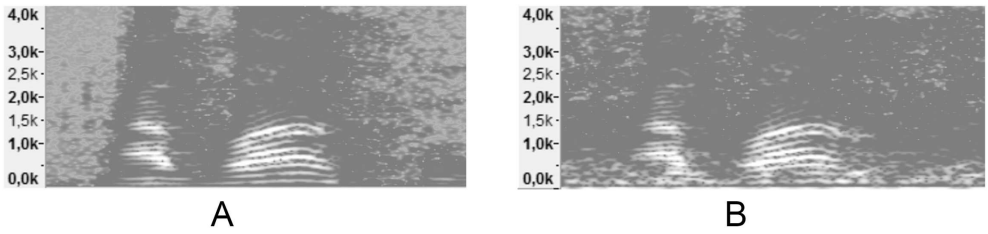

Figura 4.1: Espectogramas de áudio do comando avance (a) e com ruído aditivo de conversa (b)

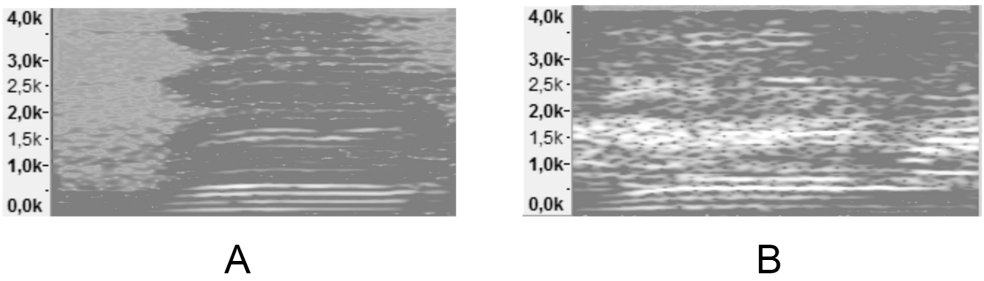

Figura 4.2: Espectogramas de áudio da palavra zero (a) e com ruído natural de conversa (b)

ruidosos, os locutores tendem a se comunicar diferentemente, aumentando o tom da voz, forma de pronúncia e duração da sílabas e palavras, etc. Essa tendência é conhecida como efeito Lombardi (Junqua, 1996). As Figuras 4.1 e 4.2 apresentam espectogramas de áudio com e sem ruídos. O exemplo apresentado na Figura 4.2, com ruído natural, mostra forte alteração em concentrações de energia, se comparado com o exemplo da Figura 4.1, com ruído aditivo.

Após essa constatação, um segundo experimento foi idealizado. Nele é usada uma segunda base, descrita na Seção 4.2, que possui áudios com três ruídos simulados durante a gravação. Desta forma, pretende-se avaliar o modelo CNN e os classificadores SVM e GMM em condição de independência de locutor e ruído real.

4.1.1 Treinamento

O processo de treinamento dos modelos é estruturado conforme diagrama da Figura 4.3.

A primeira etapa do treinamento consistiu em realizar anotações fonéticas nas bases de áudio. Foi realizada uma anotação fonética semiautomática a partir do plugin EasyAlign para o Praat (Goldman, 2011). A anotação é semiautomática pois o EasyAlign acaba confundido as marcações de fonemas na maioria dos áudio, o que exige uma supervisão humana para alinhar com mais precisão. Na Figura 4.4, é apresentada a in-

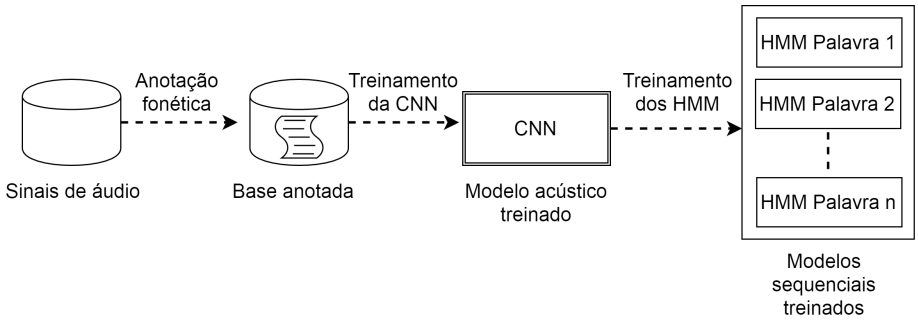

Figura 4.3: Diagrama do processo de treinamento

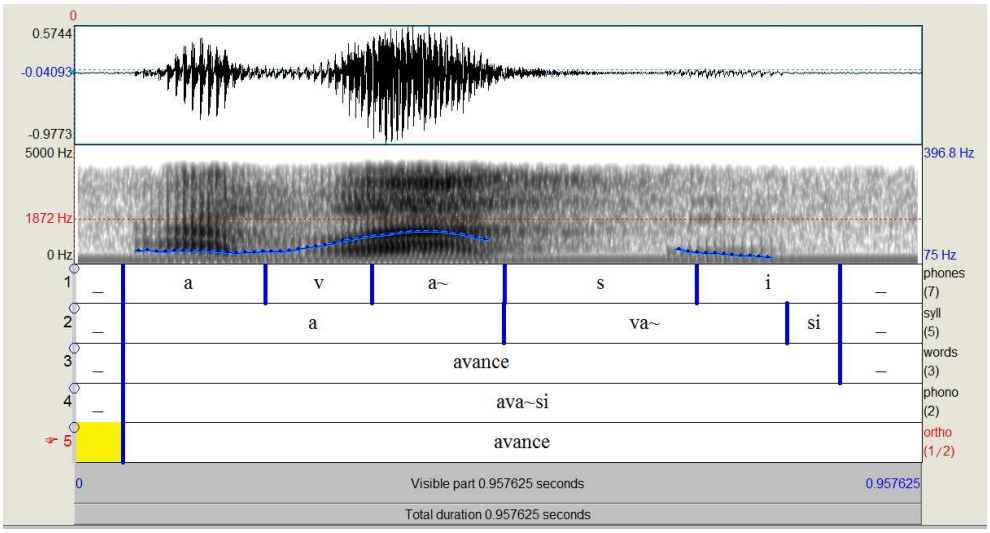

Figura 4.4: Tela do praat, utilizado para a auxiliar na anotação fonética

terface do Praat com a anotação feita para a palavra avance. Cada fonema é delimitado e, em seguida, anotado num arquivo para ser usado pelo scripts de treinamento.

Logo em seguida é realizado o treinamento da CNN. A CNN foi implementada na linguagem Python 3.4 com o auxílio das bibliotecas Numpy 1.8.1 (Jones et al., 01) e Theano 0.6 (Theano Development Team, 2016). O treinamento consiste em fornecer frames de áudio das bases e os rótulos obtidos com a anotação fonética. A abordagem proposta não usa modelo de rejeição na classificação. A CNN treinada serve para gerar as probabilidades que serão usadas no treinamento dos modelos HMM usando o algoritmo de Baum–Welch (Rabiner, 1989). Para cada palavra, é treinado um HMM que irá aprender a estrutura sequencial da palavra. Foram utilizados apenas modelos sem ruídos aditivos ou naturais.

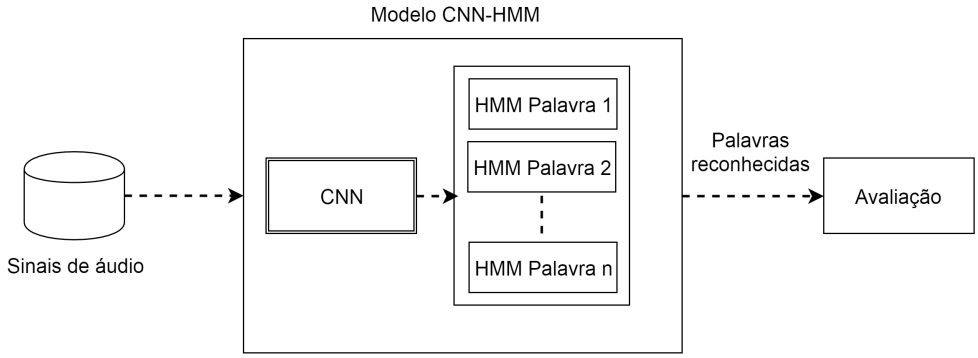

Figura 4.5: Diagrama do processo de teste

4.1.2 Teste

A etapa de teste, representada na Figura 4.5, utiliza os modelos construídos para fazer a avaliação da abordagem proposta.

4.2 Base de dados

O primeiro experimento foi realizado na base Biochaves [1] que é formado por cinco comandos pronunciados em português brasileiro (“avance", “direita", “esquerda", “pare"e “recue"). Os comandos são repetidos 10 vezes por 8 locutores (6 homens e 2 mulheres). As gravações foram realizadas em ambiente não controlados, a partir de dispostivos móveis a uma taxa de 8 Khz e quantização de 16 bits(Raulino et al., 2013). A base foi anotada foneticamente para este trabalho através de um processo semiautomático usando o plugin EasyAlign para o Praat (Goldman, 2011). Foram usados os 15 fonemas que formam os comandos, além de mais uma classe que representa *frames* de silêncio.

Os resultados foram comparados com Almeida (2014), no experimento que envolve o reconhecimento de palavras isoladas com ruídos aditivos: conversa, volvo e fábrica. Para a realização do experimento, a base foi dividida em duas bases de treinamento e de teste, com 70% das amostras para treinamento e o restante para teste. Os ruídos aditivos foram aplicados na base limpa mantendo uma relação sinal-ruído, SNR, do inglês *signal-to-noise ratio* de 6dB. Seguindo os mesmos critérios aplicados por Almeida (2014), a rede foi treinada com todos os locutores, caracterizando o problema como dependente de locutor. Em seguida, a base de teste foi modificada pelos ruídos aditivos, gerando assim, outras três bases.

A segunda base utilizada no segundo experimento foi construída pelo grupo de pesquisa do qual o autor deste trabalho faz parte. A base é formada por dígitos de zero a nove e gravada por 13 locutores, sendo 6 homens e 7 mulheres. Foram usadas as mesmas configurações de gravação da base Biochaves e anotação fonética semiautomática com

[1] http://www.biochaves.com/en/download.htm

Tabela 4.1: SNR_{dB} por locutor da base numérica

Locutores	SNR_{dB}		
	Chuva	Conversa	Rua
Homem 1	6,37dB	-0,26dB	-1,02dB
Homem 2	3,36dB	1,68dB	1,43dB
Mulher 1	1,64dB	-1,45dB	0,15dB
Mulher 2	5,43dB	2,94dB	0,89dB

o Praat. Foram escolhidos 18 fonemas e um símbolo de silêncio, formando um total de 19 rótulos para a classificação.

Quatro locutores, 2 homens e 2 mulheres, repetiram suas gravações outras 3 vezes para a base de teste, com um ruído de ambiente diferente em cada gravação. Os ruídos utilizados foram: conversa, chuva e rua. Esses quatros locutores foram utilizados para teste, com o objetivo de avaliar a proposta de forma independente de locutor. Os sinais com ruídos mantiveram diferentes SNR para cada locutor e ruído, como pode ser visto na Tabela 4.1.

Para cada elocução, uma etapa de extração de características produziu vetores de 40 coeficientes de banco de filtros na escala Mel. Abdel-Hamid et al. (2014) afirma que MFCC convencional não é adequado para ser usado na CNN porque a transforma discreta do cosseno projeta a energia espectral para uma nova base que pode não preservar informações espaciais nos áudios. A extração foi realizada em frames de 25 milissegundos com 10 milissegundos de entrelaçamento entre eles. Os coeficientes foram agrupados em 15 frames consecutivos para servirem como entrada na rede. A entrada foi rotulada de acordo com o frame da oitava posição.

4.3 Métricas

Em Almeida (2014), foram usados vetores de características num processo de alinhamento temporal conhecido como *Dynamic Time Warping* (DTW). Portanto, foi mais apropriado em seu trabalho, analisar o problema como um problema de detecção, ao invés de um problema de classificação. Desta forma, a métrica *Equal Error Rate* (EER) foi usada como forma de representar os resultados na comparação com Almeida (2014). EER corresponde ao ponto na curva ROC onde a taxa de falsos positivos e falsos negativos são iguais. Outra métrica usada foi a acurácia da classificação, aplicada tanto no primeiro como no segundo experimento. A acurácia é definido como a proporção entre a quantidade de acertos e a quantidade total de amostras na base de teste. Na avaliação da acurácia, os modelos foram submetidos a um processo de validação cruzada *k-fold* com $k = 10$.

Os extratores comparados em Almeida (2014) foram MFCC, ZCPA (*Zero-Crossing with Peak Amplitudes*) e eventos acústicos elementares (*Elementary Acoustic Events*, EAE), que foi proposto pela autora.

No primeiro experimento deste trabalho, considerou-se que a CNN está agindo no

papel de extrator de características, avaliando uma possível representação do espectograma natural como entrada, os coeficientes Fbank. Entretanto, é importante ressaltar que o modelo GMM-HMM tradicional (Rabiner, 1989) assim como o CNN-HMM deste trabalho, realizam extração de características e detecção de sinal de forma conjunta. Além disso, para o CNN-HMM proposto, foi necessário usar uma base anotada foneticamente, em contraste com o trabalho de Almeida (2014), que não dispõe de anotação fonética, apenas de anotação da elocução.

4.4 Resultados

4.4.1 Primeiro Experimento

No primeiro experimento, a CNN foi avaliada como um extrator de características para realizar a comparação de resultados com o EAE e demais características na base Biochaves. Desta forma, o reconhecimento dos comandos foi feito através de um processo de alinhamento temporal entre duas matrizes de probabilidades geradas pela CNN. A primeira matriz serve como sinal de referência, enquanto que a segunda é um sinal desconhecido que pode ser ou não da mesma classe. O problema passa a ser um problema de detecção, no qual se pode extrair medidas de EER para serem comparadas aos resultados de Almeida (2014). O método de alinhamento da referência usado foi o DTW, que mede a similaridade entre duas sequências temporais (Itakura, 1975). Na Tabela 4.2, são apresentados os resultados dos experimentos da CNN e DTW. Todos os métodos foram avaliados nas bases modificadas e não modificadas.

O modelo CNN-DTW obteve os melhores resultados em todas as bases, quando comparados ao EAE em Almeida (2014). É importante notar que o CNN-DTW foi o único treinado com ajuda da anotação fonética da base de áudios.

Os resultados da Tabela 4.2 sugerem que anotação semiautomática dos fonemas teve grande influência na melhoria de desempenho, por dispor da supervisão humana durante a fase de treinamento. Portanto, dado que os classificadores GMM e SVM (Support Vector Machines) podem ser treinados com anotação, ambos classificadores foram avaliados no mesmo experimento para averiguar a influência da anotação no processo. Várias configurações foram testadas para GMM em busca do número de Gaussianas que fornece o melhor desempenho conforme apontada na descrição da tabela de resultados 4.3. Para o SVM, um kernel polinomial de grau $d = 3$ alcançou os

Tabela 4.2: Resultados do primeiro experimento com CNN DTW

	Equal error rate (EER) (%)			
Métodos	Base limpa	Conversa	Volvo	Fábrica
MFCC-DTW	7.30%	11.70%	8.20%	13.10%
ZCPA-DTW	8.20%	10.50%	8.40%	12.20%
EAE-DTW	3.80%	8.50%	3.80%	9.90%
CNN-DTW	**3,01%**	**5,96%**	**3,42%**	**6,97%**

Tabela 4.3: EER do primeiro experimento com anotação fonética

Métodos	Equal error rate (EER) (%)			
	Base limpa	Conversa	Volvo	Fábrica
GMM-HMM (1)	19,20%	28,27%	19,23%	23,20%
GMM-HMM (3)	4,27%	27,93%	4,20%	20,20%
GMM-HMM (5)	1,07%	29,33%	1,07%	14,67%
SVM-HMM	**0,00%**	20,27%	**0,00%**	15,13 %
CNN-HMM	0,13%	**3.47%**	0,27%	**2,67%**

melhores resultados. Os classificadores foram implementados pelo projeto scikit-learn[2], uma biblioteca de código aberto com implementações de algoritmos para aprendizado de máquina na linguagem Python (Pedregosa et al., 2011). Os dois classificadores geram sequências de observações para o treinamento de HMM discreto.

A Tabela 4.3 mostra os resultados da CNN comparados com GMM e SVM. O número ao lado dos GMMs indica a quantidade de Gaussianas na mistura. O SVM obteve resultados um pouco melhores na base limpa e com ruído "volvo", mas o CNN continuou com melhores resultados nas demais situações. Os modelos também foram avaliados usando validação cruzada k-*fold* da acurácia com $k = 10$. Na Tabela 4.4 são apresentadas as médias da acurácia de cada teste seguidas pelo desvio padrão para cada situação.

Tabela 4.4: Acurácia do primeiro experimento com anotação fonética

Métodos	Acurácia (%)			
	Base limpa	Conversa	Volvo	Fábrica
GMM-HMM (1)	79,20% ± 2,14	57,80% ± 7,39	78,40% ± 1,95	71,40% ± 3,40
GMM-HMM (3)	94,50% ± 1,17	63,80% ± 11,33	94,50% ± 1,17	76,80% ± 3,42
GMM-HMM (5)	99,80% ± 0,42	64,20% ± 13,18	99,80% ± 0,42	82,40% ± 4,29
SVM-HMM	**100,00%**	77,40% ± 4,22	**100,00%**	83,20 % ± 3,29
CNN-HMM	99,67% ± 1,02	**88,91%** ± 5,43	99,67% ± 1,02	**94,86%** ± 4,53

SVM e GMM tiveram seus desempenhos afetados pelas bases modificadas com ruídos de conversa e de fábrica. Por serem ruídos mais espalhados em diversas faixas de frequências, todas as máquinas de aprendizado foram afetadas. Contudo, foi possível notar que CNN-HMM foi menos influenciado por tais ruídos, indicando desta forma que o modelo tem um nível maior de robustez a variações comuns em ambientes não controlados. A Figura 4.6 apresenta a curva ROC para o CNN-HMM com ruído de conversa.

[2] http://scikit-learn.org

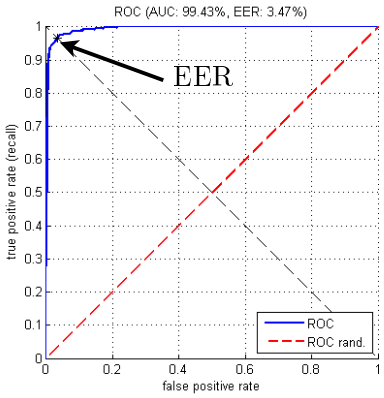

Figura 4.6: Curva ROC para a base modificada com ruído de conversa

4.4.2 Segundo Experimento

A Tabela 4.5, apresenta os resultados do segundo experimento realizado na base numérica. O desempenho do CNN-HMM foi comparado novamente com SVM e GMM. Para o GMM, apenas foi considerada a configuração com cinco gaussianas que havia apresentado melhor desempenho anteriormente. Os modelos foram avaliados também num contexto de reconhecimento de fala dependente de locutor e seus desempenhos são apresentados na Tabela 4.6.

Tabela 4.5: Resultados do segundo experimento independente de locutor

	Acurácia (%)			
Métodos	Base limpa	Chuva	Conversa	Rua
GMM-HMM (5)	40,00% ± 7,71	10,00%	9,50% ± 1,00	11,50% ± 1,91
SVM-HMM	44,40% ± 8,98	18,50% ± 6,40	16,00% ± 9,52	10,50% ± 1,00
CNN-HMM	**59,79%** ± 13,54	**32,50%** ± 7,00	**34,00%** ± 10,19	**23,00%** ± 4,76

Tabela 4.6: Resultados do segundo experimento dependente de locutor

	Acurácia (%)			
Métodos	Base limpa	Chuva	Conversa	Rua
GMM-HMM (5)	95,80% ± 4,04	12,50% ± 3,00	9,50% ± 1,00	12,00% ± 1,63
SVM-HMM	99,80% ± 0,63	18,50% ± 7,72	12,00% ± 4,00	15,00 % ± 2,00
CNN-HMM	**100,00%**	**52,00%** ± 4,61	**30,50%** ± 7,54	**18,00%** ± 3,26

Os resultados apontam novamente o CNN-HMM como a melhor alternativa entre os modelos testados. Apesar de ter sido menos afetado que os demais, o CNN-HMM obteve fraco desempenho nessa segunda base, principalmente quando se trata do ruído de rua que é caracterizado como mais intenso entre os três. Isso pode ter sido causado por se

tratar de uma base com pouca quantidade de amostras, dificultando a generalização do aprendizado da CNN, crucial no reconhecimento de um padrão tão diversificado como a fala. Outro ponto importante foi a utilização de ruídos muito intensos durante a coleta de dados como pode ser visto na Seção 4.2. As distorções causadas por esses ruídos não foram filtradas como muitas vezes ocorre durante a captura de áudio, o que pode ter levado ao desempenho inferior ao experimento anterior. Os modelos obtiveram melhores resultados no reconhecimento dependente de locutor, conforme Tabela 4.6, mas ainda assim não conseguem manter o desempenho nas bases afetadas por ruídos.

Capítulo 5

Aplicações

O modelo CNN-HMM treinado a partir dos comandos da base Biochaves foi utilizado para o desenvolvimento de dois jogos pelo grupo de pesquisa de reconhecimento de fala. Os jogos foram desenvolvidos na versão gratuita do motor de jogos Unity3D. Essa ferramenta é um editor poderoso que permite criação dos mesmos em 2D e 3D com física própria, aplicando-a para simular as leis físicas atuante no mundo real, grande documentação e tutoriais disponíveis.

Os scripts podem ser feitos em C# ou Javascript, sendo escolhido neste projeto o C# por haver maior documentação e tutoriais disponíveis. Desenvolver com unity ainda permite a facilidade de ter projetos multiplataforma como Android, IOS, Windows Phone, Windows, Mac OSX, Linux, Samsung TV, Playstation, Xbox etc. Além disso, pode usar os serviços integrados de Unity para acelerar seu processo de desenvolvimento e otimizar seu jogo como recursos de monetização, *networking* para *multiplayer*, *cloud building*, *asset store* e demais. A *asset store* é uma loja da Unity com vários recursos gratuitos e pagos para desenvolvimento. Esses recursos são modelos 3D, animações, áudios, modelos de projetos, *spritesheet*, texturas, pacotes de efeitos de câmeras, *plugins* que permitem extensibilidade da *engine* como para o sensor de movimentos Kinect.

Um jogo é baseado em sistema de turnos do RPG, e o outro requer resposta em tempo real. Esses jogos foram feitos para analisar o impacto desse tipo de reconhecedor em ambas as situações.

5.1 Jogo Cálculo de Aventura

O jogo Cálculo de Aventura é um jogo de aventura com temática medieval e possui música ambientada. Tanto as músicas como os *sprites* foram baixados na *asset store*. As Figuras 5.1 e 5.2 apresentam telas do Cálculo de Aventura. A prototipação do jogo seguiu os seguintes passos:

- O *game design*: conceitualização da história do jogo, personagens, vilões, interfaces do usuário (fácil cognição que auxilie na execução do comando), público alvo,

Figura 5.1: Tela inicial

Figura 5.2: Exemplo de fase do jogo

estilo de cores, ambientação, jogabilidade, regras do jogo, controles do jogo (botão *touch* para envio do áudio), esboços de telas;

- Escolha da *engine*;
- Buscar documentação da *engine* e jogos relacionados com reconhecedor de fala;
- Implementação;
- Testes de usabilidade.

Esse jogo que está em desenvolvimento é um RPG *single player* 2D de tabuada para crianças entre 7 a 9 anos. Ele possui 6 fases, sendo a última o *boss* final da história. O personagem é um guerreiro que busca encontrar a princesa desaparecida e salvar o vale do ataque de um Dragão, Tiamática. O dragão é controlado pelo Mestre dos Cálculos que passa os desafios para o jogador. Desde o ataque do Dragão, a princesa havia sumido.

Em cada fase a criança deve acertar o desafio do mestre para ganhar *level*, derrotar os dragões menores e seguir seu destino até o castelo. O desafio do mestre consiste em calcular algumas operações matemáticas de soma, subtração, multiplicação e divisão de dois ou três termos positivos de 2 algarismos, a depender da dificuldade . O jogo apresenta quatro opções de resposta, sendo apenas uma verdadeira. Não é permitido chutar, pois se acertar ou errar, sempre é gerada uma nova pergunta. Cada opção de resposta gera uma ação no personagem, sendo essas ações: "ataque com espada" quando falar avance, "defesa com escudo" quando falar recue, 'usar poção de cura" quando falar direita e "usar magia de fogo" quando falar esquerda.

Para responder os desafios, o jogador deve apertar no botão do microfone, que fará o cliente ficar aberto ouvindo por 2 segundos. Após os 2 segundos, ele fecha o microfone, inicia e estabelece conexão com o servidor, envia o áudio e aguarda a reposta da palavra reconhecida. Em seguida, é computado se a resposta foi certa ou errada, além de executar a ação mapeada para aquela palavra. Se a resposta for errada, a ação do personagem é executada, porém ocorrerá o "*miss*", dando ataque de oportunidade do inimigo. Ataque de oportunidade é conhecido no RPG como sendo uma brecha para ataque do inimigo que antecipa sua ação.

A proposta da execução de uma ação vem para fazer a criança interagir mais no mundo do jogo, onde ela deve pensar qual seria a ação "necessária" para combater aquele dragão. Cada operação matemática tem 30 segundos para ser respondida, antes de gerar nova pergunta. O tempo estipulado foi feito levando em conta o tempo de 5 segundos para reconhecimento da palavra e o tempo observado que uma criança de 7 anos demora para fazer as contas usando os dedos das mãos.

Cada fase exige um número de acertos para seguir para a próxima fase, e o jogador não é penalizado com dano ou perda de pontos quando erra, apenas deixa de ganhar *level*. O *boss* final, Tiamática, é um dragão de quatro cabeças representando as quatro operações básicas. Nele, o jogador deverá responder contas aleatórias dentre as 4 operações, e para cada uma solucionada, ele retira vida do vilão. Num momento crítico de

vida do inimigo, o jogador tem a opção de matar o dragão, ou nocautear. Ressaltando que por se tratar de um RPG, cada ação tem uma consequência, e caso escolha matar, terá salvo o reino mas nunca terá notícias sobre a princesa desaparecida. Se escolher nocautear, será revelada e desfeita a maldição sobre o dragão, onde ele se transformará na princesa desaparecida. O encanto só poderia ser quebrado quando um guerreiro corajoso e justo derrotasse o dragão sem matá-lo, e é preciso fazer a escolhe certa para descobrir ou jogar novamente.

O *game design* foi feito pensando numa proposta lúdica e educativa, sendo a segunda baseada no comportamento operante do psicólogo Skinner que conceitualiza o reforço, punição e extinção de comportamentos para aprendizado de uma tarefa. Esse jogo requer que o jogador já tenha conhecimentos de tabuada, e serve para a prática da mesma de forma interativa.

5.2 Jogo Breaker Aracaju

O jogo é baseado em um clássico dos games, cujo nome é Break Ball. No entanto, este é nomeado como Break Aracaju, igualmente, houve uma personalização nele, no lugar da bola foi usado um caju, fruta típica de Aracaju-SE. A utilização de comandos é feita através da voz para controlar a plataforma, pois o objetivo desse jogo é não deixar o caju cair no chão para isto não acontecer, precisa-se controlar a plataforma feita de folha, ela deve-se mover para direita, esquerda ou parar de acordo com o caju. Simultaneamente deve-se destruir os blocos de madeiras acima da plataforma, para conseguir maior pontuação e vencer. O jogo foi pensado e construindo tendo no momento uma fase. A Figura 5.3 apresenta a tela inicial e a Figura 5.4 a tela padrão do jogo.

Figura 5.3: Tela inicial

O botão verde com simbolo do microfone, ao ser apertado, ativa o cliente do jogo que ativa a captação do áudio pelo microfone da plataforma na qual o jogo esta rodando e captura o áudio falado com duração de dois segundos. Uma vez processado o sinal no servidor e reconhecido de acordo com as palavras que o jogo utiliza (Direita,

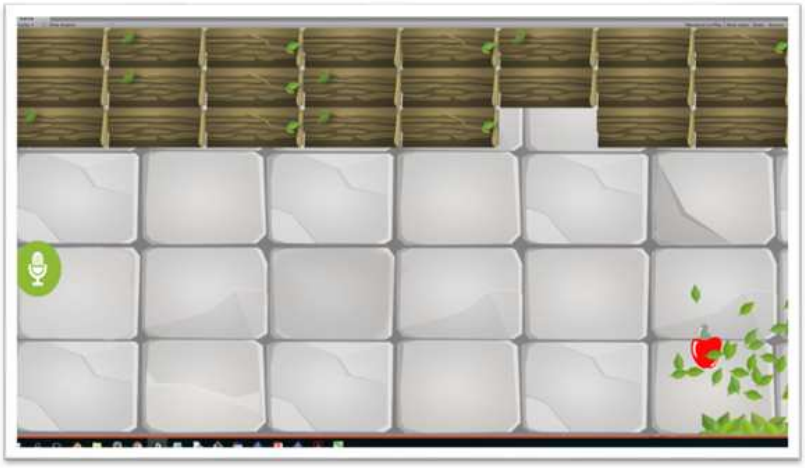

Figura 5.4: Exemplo de fase do jogo

Esquerda e Pare). De acordo com cada palavra a plataforma se movimenta. Ao final do jogo, mostra-se a pontuação, sendo esta baseada no seguinte calculo, número de blocos destruídos dividido pelo número total de blocos existentes.

Capítulo 6

Conclusão

Este trabalho apresentou um modelo híbrido entre CNN e HMM para reconhecimento de fala

Os resultados dos experimentos mostraram que o modelo CNN-HMM consegue melhorar o reconhecimento de palavras mesmo na presença de ruído. Mesmo sendo um modelo que não usa nenhum método adicional para o tratamento do ruído, a CNN é capaz de normalizar algumas variações acústicas que podem ocorrer em ambiente não controlados. Essa melhoria pode ter sido influenciada pela anotação fonética da base, contudo, como visto nos resultados, a pequena variação entre as taxas de EER encontradas entre as bases mostram a robustez da abordagem.

Outra propriedade importante do modelo é a capacidade de usar a rede como *front-end* do sistema ASR. Através das camadas de convolução e subamostragem, a CNN consegue realizar uma extração implícita de características dos dados de entrada. Além disso, propriedades como compartilhamento de pesos e conectividade local introduzem um certo nível de invariância a distorções presentes em situações reais. As camadas de subamostragem são capazes de tratar pequenos deslocamentos no domínio da frequência que são bem comuns em sinais de fala.

O modelo, nesta abordagem, possui a desvantagem de necessitar de uma anotação fonética das palavras para que possa ser treinado. Esse tipo de anotação nem sempre está disponível em bases de treinamento, e demandam de tempo para serem realizadas. Além disso, alguns fonemas são bastante curtos, podendo criar confusão durante o trabalho de anotação e consequentemente alguns fonemas podem ter sua rotulação equivocada.

Como extensão deste trabalho, pretende-se realizar os seguintes trabalhos:

- Adaptação do modelo para o reconhecimento de fala contínua em amplo vocabulário;
- Analisar e propor métodos para realizar a anotação fonética de bases de fala;
- Experimentos com outros tipos de problemas, como: reconhecimento de locutor, gestos, etc.

Referências

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(10):1533–1545.

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4277–4280. IEEE.

Almeida, C. R. (2014). Extratores de características acústicas inspirados no sistema periférico auditivo. Master's thesis, Federal University of Sergipe, São Cristóvão.

Bourlard, H. A. and Morgan, N. (1994). *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media.

Chang, S.-Y. and Morgan, N. (2014). Robust cnn-based speech recognition with gabor filter kernels. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE.

Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., and Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 1237.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

Goldman, J.-P. (2011). Easyalign: an automatic phonetic alignment tool under praat. Interspeech'11, 12th Annual Conference of the International Speech Communication Association.

Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Huang, J.-T., Li, J., and Gong, Y. (2015). An analysis of convolutional neural networks for speech recognition. In *ICASSP*.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):67–72.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2016-05-17].

Junqua, J.-C. (1996). The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, 20(1):13–22.

Jurafsky, D. and Martin, J. H. (2000). *Speech & language processing*. Pearson Education India.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., Huang, F. J., and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–97. IEEE.

Lee, H., Pham, P., Largman, Y., and Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(4):745–777.

Mitra, V., Wang, W., Franco, H., Lei, Y., Bartels, C., and Graciarena, M. (2014). Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raulino, C., Duarte, D., and Montalvão, J. (2013). Análise de espectro através da detecção de eventos acústicos elementares no plano tempo-frequência. In *Simpósio Brasileiro de Automação Inteligente*.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426.

Smith, A. (2015). U.s. smartphone use in 2015. *Pew Research Center*.

Soltau, H., Kuo, H.-K., Mangu, L., Saon, G., and Beran, T. (2013). Neural network acoustic models for the darpa rats program. In *INTERSPEECH*, pages 3092–3096.

Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.

Varga, A. and Steeneken, H. J. (1993). Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251.