Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Investigação do esforço necessário nas etapas do treinamento de modelos acústicos e de linguagem para transcrição de áudio e seu impacto na acurácia de modelos

Marcos Valadão Gualberto Ferreira

JUIZ DE FORA
NOVEMBRO, 2017

Investigação do esforço necessário nas etapas do treinamento de modelos acústicos e de linguagem para transcrição de áudio e seu impacto na acurácia de modelos

Marcos Valadão Gualberto Ferreira

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Jairo Francisco de Souza

JUIZ DE FORA

NOVEMBRO, 2017

Investigação do esforço necessário nas etapas do treinamento de modelos acústicos e de linguagem para transcrição de áudio e seu impacto na acurácia de modelos

Marcos Valadão Gualberto Ferreira

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Jairo Francisco de Souza
Dr. em Informática (PUC-RIO)

Eduardo Barrére
Dr. em Engenharia de Sistemas e Computação (COPPE/UFRJ)

Fabricio Martins Mendonça
Dr. em Ciência da Informação (UFMG)

JUIZ DE FORA

30 DE NOVEMBRO, 2017

À todos que fizeram parte desta conquista

Resumo

Reconhecimento de fala é um tema recorrente nas áreas de Recuperação de Informação e web. A possibilidade do computador processar sinais de áudio e gerar transcrições textuais criam uma série de aplicações para estas informações. Os sistemas de reconhecimento de fala utilizam modelos estatísticos, que são construídos através de treinamento supervisionado, e regras de estruturas de linguagem. O maior desafio desses sistemas é treinar os modelos acústicos e de linguagem com o objetivo de maximizar a acurácia do texto transcrito . Este treinamento é um processo caro pois necessita de uma base de arquivos consideravelmente grande e bem processada e demanda o cumprimento de diversas tarefas, tudo isto para encontrar uma modelagem satisfatória. A abordagem proposta neste trabalho tem como objetivo explorar as tarefas pertinentes no treinamento de modelos acústicos e de linguagem com o objetivo de encontrar quais delas mais influenciam na acurácia final do modelo e direcionar melhor o tempo, empenho e desenvolvimento de melhorias. Através dos experimentos realizados neste trabalho, conclui-se que o investimento no processamento da base utilizada no treinamento de modelos resultou no melhor ganho, com relação a WER (Word Error Rate), e possibilitou a criação de um sistema de reconhecimento de fala robusto e com possibilidades de aplicações.

Palavras-chave: Reconhecimento de fala, treinamento de modelos, modelos acústicos, modelo de linguagem.

Abstract

Speech recognition is a recurring theme in the areas of Information Retrieval and web. The ability of the computer to process audio signals and generate textual transcriptions creates a series of applications for this information. Speech recognition systems use statistical models, which are constructed through supervised training, and rules of language structures. The greatest challenge of these systems is to train the acoustic and language models with the aim of maximizing the accuracy of the transcribed text. This training is an expensive process because it requires a considerably large and well-processed file base and demands the fulfillment of various tasks, all to find a satisfactory modeling. The approach proposed in this work aims to explore the pertinent tasks in the training of acoustic and language models with the objective of finding which ones more influence the final accuracy of the model and to better target the time, commitment and development of improvements. It was concluded that the investment in the processing of the base used in the training of models resulted in the best gain, in relation to WER (Word Error Rate), and enabled the creation of a robust speech recognition system and with possibilities of applications.

Keywords: Speech recognition, training models, acustic model, language model.

Agradecimentos

Agradeço a Deus por meus pais, Carlos e Ivanete, que dedicaram tudo na educação minha e de meus irmãos. Obrigado pelo apoio incondicional, pelo sustento, pela motivação e principalmente pelo amor que foi essencial em toda minha formação. Vocês são meus referenciais de vida e fé em Deus.

Agradeço a Deus por minha namorada, Débora, que tem sido minha companheira durante todo o curso, por ter escutado minhas reclamações e acalmado meu coração, por ter dedicado amor e alegria durante este tempo, pelos conselhos e por todos os momentos que passamos juntos. Você é o amor da minha vida.

Agradeço a Deus por minha grande família, Sandro, Márcia, Beatriz, Tiago, Tiffany, Yasmin, Jade, Lucas, Karine, Maria Luiza, Pedro Lucas, Danielle, Mateus, Maria Eduarda, Roberta e Aloísio, por alegrarem a minha vida, por terem cuidado de mim em Juiz de Fora, me apoiarem e se alegrem com minhas conquistas, na verdade, nossas conquistas. Vocês são meus exemplos de vida.

Agradeço a Deus por todos que foram presentes nesta caminhada, ao meu sogro, cunhado e a toda família da Débora que me apoiaram durante o curso e aos meus demais familiares.

Agradeço a Deus pelos inúmeros amigos que fiz durante o curso, em especial, a galera do Lapic, por me apoiaram nos últimos dois anos de estágio e ao José Eduardo pelo apoio, paciência e pelos ensinamentos. Aos demais amigos que fiz em Juiz de Fora, que se tornaram verdadeiros irmãos. Vocês são meus companheiros.

Agradeço a Deus pelos professores do Departamento de Ciência da Computação e por seus ensinamentos. Em especial, aos mestres Jairo e Eduardo Barrére, que tanto me ensinaram e motivaram a concluir o curso e a ser um profissional melhor.

Enfim, agradeço a Deus por, além de ter colocado pessoas especiais em minha vida, guardou-a, foi meu sustento, me ensinou, e esteve presente em cada um desses momentos. Você é o bem mais precioso que possuo.

"Ter-se a consciência de que se é ignorante, constitui um grande passo na direção da sabedoria." *Benjamin Disraeli*

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

ASR	*Automatic Speech Recognition*
CTS	*Conversation Telephone Speech*
DCC	Departamento de Ciência da Computação
DNN	*Deep Neural Networks*
GMM	*Gaussian mixture model*
HMM	*Hidden Markov Model*
LM	*Language Model*
LPCC	*Linear Predictive Cepstral Coefficients*
MFCC	*Mel-Frequency Cepstral Coefficients*
OOV	*Out of Vocabulary*
PB	Português Brasileiro
PLN	Processamento de Linguagem Natural
UFJF	Universidade Federal de Juiz de Fora
WER	*Word Error Rate*

1 Introdução

Um sistema de Processamento de Linguagem Natural (PLN) abrange aspectos da comunicação humana como a fala, palavras, textos e sentenças considerando todo o contexto em que se encontra e se baseando em estruturas de linguagem (Gonzalez e Lima, 2003). Estes sistemas são utilizados, de maneira geral, para possibilitar que o computador entenda e se comunique em linguagem humana de forma automática. O PLN é baseado em modelagens realistas que sempre necessitam de melhorias por parte das estruturas de comportamento de linguagem (Gonzalez e Lima, 2003).

Existem diversas aplicações na área de PLN, entre elas está a técnica de reconhecimento automático de fala (ou ASR, de *automatic speech recognition*), que é alvo de estudo deste trabalho.

Tecnologias de ASR permitem que computadores interpretem a fala humana a partir da síntese de sinais de áudio, ou seja, transcrever áudio em textos de linguagem natural (Coelho e Souza, 2015). Este processo geralmente apresenta grandes dificuldades pois possui limitações quanto ao reconhecimento por questões técnicas tais como: tamanho do vocabulário, reconhecimento de fala contínua e quanto a estrutura complexa da voz humana, que depende de fatores como: sotaque, entonação, velocidade da voz, estado emocional etc (Gonzalez e Lima, 2003).

Sistemas de ASR são construídos, geralmente, a partir de decodificadores e de modelagem acústica e de linguagem. Os decodificadores desempenham o papel de interpretação do sinal de áudio e de identificação, dentro da modelagem, por algum termo de equivalência, já os modelos acústicos e de linguagem são responsáveis pelo mapeamento de "toda" estrutura linguística(Cuadros, 2007). É praticamente impossível que um sistema seja perfeito a ponto de abranger as diversas formas que a voz humana pode tomar e de como interpretá-la. Por isto, os sistemas de reconhecimento de voz conhecidos não possuem acurácia de 100%.

O grande paradigma de um sistema de ASR é possuir modelos acústicos e de linguagem que sejam satisfatório (Coelho e Souza, 2015). Estes modelos são treinados

utilizando uma base de dados pré processada e técnicas de descrição probabilística para os termos da linguagem (Neto et al., 2005). Existem na literatura, diversas técnicas de treinamento de modelos acústicos e de linguagem que utilizam bases de dados de tamanho considerável, bem processadas e que geralmente são proprietárias. Entretanto, algumas técnicas paralelas ao treinamento podem criar uma base que gere um bom modelo e seja livre.

1.1 Justificativa

Uma forma de melhorar o processo de reconhecimento de fala é através de melhorias no treinamento dos modelos. Este treinamento é um processo muito custoso pois depende de diversos fatores como uma base de arquivos grande e bem processados. Essas bases geralmente são pagas e pouco especializadas. Na literatura estão presentes formas de criar bases de dados livres, como a partir de *audio books* (Panayotov et al., 2015). Entretanto, são tarefas exaustivas e demandam um processamento complexo da base como normalização, expansão de número e acrônimos, correção gramatical etc (Siravenha et al., 2008).

Como o treinamento requer a execução de diversas tarefas, uma investigação de todo o processo de treinamento, possibilitando o entendimento de quais etapas serão críticas em relação à modelagem, poderá diminuir o tempo gasto com determinadas etapas e direcionar o tempo para etapas críticas, ou seja, que mais influenciam na acurácia dos modelos.

A exploração das tarefas do treinamento, também influenciará no processamento da base de dados, indicando como alterações, melhorias e especialização do modelo de linguagem podem influenciar no resultado final e quais novos caminhos tomar em relação ao treinamento tradicional encontrado na literatura.

1.2 Objetivos

Este trabalho tem como objetivo geral verificar quais as etapas do treinamento de modelos acústicos e de linguagem mais influenciam no resultado da acurácia do modelo gerado.

Como objetivo específico, o trabalho pretende encontrar os pontos do treinamento que mais influenciam na acurácia, direcionando, desta forma, quais etapas do treinamento requerem maior esforço de tempo, processamento e supervisão. Ainda, o trabalho pretende implementar melhorias nestas etapas a fim de criar um modelo satisfatório e indicar como proceder para criar uma base gratuita, especializada e que demande menos esforço que a abordagem tradicional.

1.3 Metodologia

Para a realização do presente trabalho, serão implementados algoritmos para o treinamento de modelos acústicos e de linguagem, com o objetivo de investigar, dentre as diversas tarefas destes processos, quais influenciam na acurácia do modelo e possuem melhor custo/beneficio.

A base de áudio utilizada no treinamento de modelos acústicos será construída através de videoaulas legendadas, extraídas do repositório de cursos on-line Coursera[1]. Para a preparação da base serão utilizadas técnicas e algoritmos para segmentação do áudio e normalização de texto, para que possam estar bem alinhadas e padronizadas. A base de texto utilizada no treinamento do modelo de linguagem será extraída das legendas da base de áudio e de textos disponíveis na Web. Também serão implementados algoritmos para correção ortográfica de palavras e para padronização da base de texto. O dicionário será construído com todas as palavras da base de texto.

Para os experimentos com o treinamento de modelo acústico será utilizando o Kaldi[2] como *toolkit* de auxilio para a avaliação dos modelos gerados (Ferreira e Souza, 2017). Pretende-se no modelo acústico avaliar quais configurações de áudio e técnicas de extração de parâmetros geram o melhor resultado para o processo. Além disto, avaliaremos quais as vantagens do uso de redes neurais na geração das probabilidades dos estados.

Já para os experimentos com o treinamento de modelos de luageming, será utilizando o *toolkit* SRILM, gratuito para pesquisa, a fim de avaliar a acurácia dos modelos.

[1] `https://pt.coursera.org/`
[2] `http://kaldi-asr.org/`

Pretende-se avaliar principalmente as melhoras na base de texto e sua influência na perplexidade, que será a medida de avaliação do modelo de língua.

Espera-se com isso, criar uma base para treinamento de modelos acústicos e de linguagem que seja eficaz e gere resultados satisfatórios para a transcrição de áudio. Além disso, pretende-se concluir quais métodos e técnicas apresentam os melhores resultados para o treinamento e quais etapas são críticas e possuem o melhor custo/benefício no processo levando em consideração tempo e gastos computacionais.

2 Revisão da Literatura

A viabilidade de extrair informações de recursos multimídia como áudio possibilita uma série de aplicações em diversos contextos. Uma das aplicações mais conhecidas para reconhecimento automático de fala é o uso de transcrição simultânea na criação de legendas ocultas para sistemas de televisão. O AUDIMUS.MÉDIA (Meinedo et al., 2003) é um sistema que gera legendas ao vivo para emissoras de TV e de *streaming*. O objetivo do sistema de reconhecimento de fala, neste contexto, é gerar a legenda sem intervenção humana evitando custo com serviços de legendagem manual.

O reconhecimento de fala também é utilizado para questões de aprendizado e acessibilidade. O CineAD (Campos et al., 2014) é um sistema para geração automática de roteiros de audiodescrição (AD), que é um recurso essencial para pessoas cegas e com baixa visão para acesso ao cinema. Este trabalho utiliza, dentre outras técnicas, o reconhecimento de fala para geração de legendas que serão utilizadas na criação de AD. O reconhecimento de fala também pode ser utilizado para questões de aprendizado, como descrito em (Higgins and Raskind, 1999), onde foi avaliado o uso de um sistema ASR para auxiliar na educação de crianças e adolescentes com dificuldade de aprendizagem. O reconhecimento de fala também pode ser usado, na área da educação, para melhorar o desempenho de crianças e adolescentes como descrito em (Hämäläinen et al., 2013), onde é apresentado um jogo educacional que utiliza a técnica de reconhecimento de fala com o objetivo de melhorar a coordenação física e as habilidades em matemática e música dos estudantes.

Ainda, podemos citar aplicações práticas e cotidianas para sistemas de reconhecimento de fala. O uso de assistentes virtuais em smartphones está presente no sistema das principais empresas da área de telefonia móvel. Os assistentes são elaboradas para permitir que o usuário interaja com o dispositivo a partir da voz, utilizando assim um sistema de reconhecimento de fala e técnicas para processamento do resultado do reconhecimento. Estes sistemas de reconhecimento também podem ser encontrados em diversos dispositivos que interagem com usuários, como por exemplo carros, eletrodomésticos, sistemas de

autenticação, casas inteligentes etc. No âmbito dos negócios, o CALO (Tur et al., 2008) é um assistente de reunião, que fornece informações automáticas de reuniões utilizando, dentre outras técnicas, o uso de reconhecimento de fala para legendagem, geração de ata e busca por conteúdos de conversas.

Também podemos citar aplicações diretas para sistemas de reconhecimento de fala, como os esforços de alguns trabalhos (Coelho e Souza, 2015; Yang and Meinel, 2014; Raimond and Lowis, 2012) que utilizam estes sistemas para auxiliar na anotação, busca e recomendação de vídeos. Este processo utiliza o texto de resultado da transcrição, como entrada para algoritmos de anotação semântica na intenção de identificar termos para classificar o conteúdo multimídia. Estes termos podem ser relacionados a bases de conhecimento e desta forma criar uma estrutura de dados ligados podendo assim, utilizar técnicas para recomendação de conteúdos. Assim, pode-se perceber que os cenários para aplicação de sistemas de reconhecimento de fala são amplos e que este tipo de sistema pode auxiliar tanto na automatização de tarefas quanto na extração de informações para aplicações.

Neste capitulo, será apresentado o referencial teórico em torno do trabalho proposto e a revisão da literatura no contexto de treinamento de modelos para sistemas de reconhecimento de fala.

2.1 Modelagem de sistemas de reconhecimento de fala contínua

Os sistemas de ASR atuais são baseados no reconhecimento estatístico de padrões. A Figura 2.1. abaixo é praticamente um consenso na área e é composta basicamente pelos seguintes componentes: Etapa de extração de parâmetros que converte o sinal de áudio em uma representação compacta, robusta e mais sensível ao conteúdo linguístico; modelo acústico que mapeia o sinal original que está sendo processado em palavras e sentenças; modelo de linguagem que é responsável por caracterizar a língua e condicionar a combinação de palavras, descartando frases agramaticais; dicionário fonético que representa as palavras do dicionário utilizado, com suas respectivas transcrições fonéticas; decodificador

que procura a melhor sequência de palavras num conjunto de hipóteses possíveis.

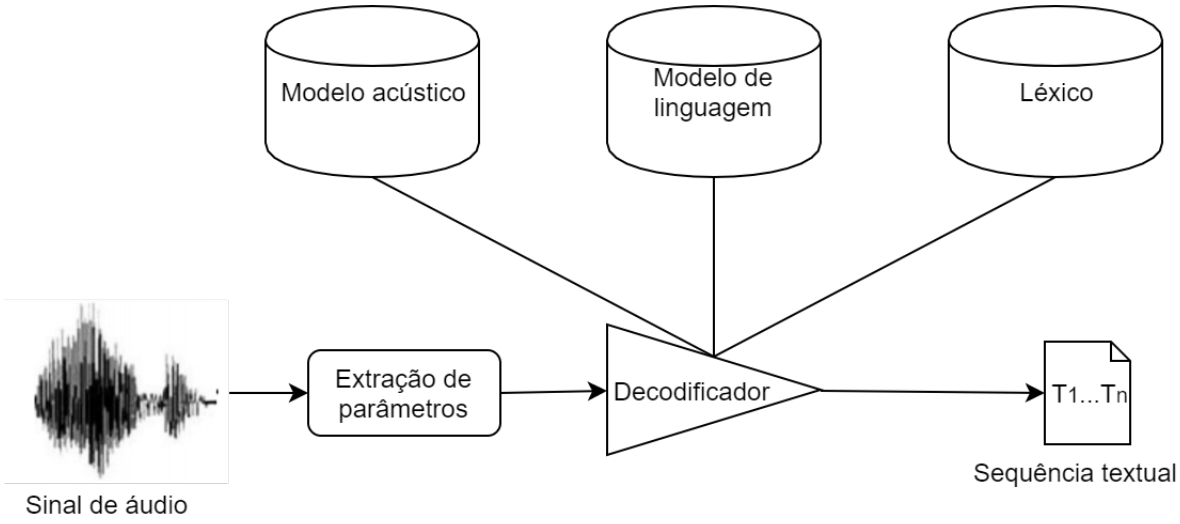

Figura 2.1: Sistema de Reconhecimento de Fala

A seguir, cada um destes componentes serão descritos com maiores detalhes.

Extração de parâmetros

A extração de parâmetros é um processo recorrente em todos os sistemas que envolvem reconhecimento de padrões. Tem como principal objetivo extrair somente as informações do áudio que serão utilizadas no reconhecimento, de forma que seja um processo objetivo e robusto (Veiga, 2013). Assim, a função prioritária deste componente é dividir o sinal que está sendo processado em blocos e de cada um destes blocos derivar uma estimativa suavizada do espectro (Tevah, 2006).

Em sistemas ASR, os parâmetros mais utilizados são coeficientes *mel-cepstrais* (MFCC – *Mel-Frequency Cepstral Coefficients*) (Davis and Mermelstein, 1980). Algumas outras abordagens de parâmetros conhecidas são: coeficientes cepstrais de predição linear (LPCC – *Linear Predictive Cepstral Coefficients*) e coeficientes dinâmicos (parâmetros delta e delta-delta) que também podem ser usados em conjunto com as demais abordagens.

Um exemplo de aplicação da extração de parâmetros em sistemas ASR é para compensar o efeito de longa duração de espectros, causado principalmente pela forma acústica como o áudio foi construído (equipamentos, canais de áudio, distância entre o locutor e o microfone etc), neste caso é utilizada uma técnica de atualização da média

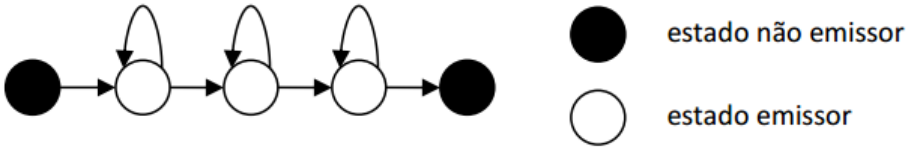

Figura 2.2: Modelo HMM com topologia esquerda-direita

espectral para cada segmento de forma a proteger o sistema contra variações do canal (Tevah, 2006).

Modelagem Acústica

Os modelos acústicos geralmente são HMM (*Hidden Markov Model*) (Rabiner e Juang, 1986) de fonemas com 3 estados emissores. Podem ser modelos sem contexto, monofones, ou com contexto: difones (contexto à direita ou à esquerda) trifones (contexto à direita e à esquerda) ou n-fones.

Um HMM é definido pelos seguintes parâmetros: número de estados, matriz de probabilidade de função entre estados e, para cada estado, uma função de densidade de probabilidade, que tem como objetivo a caracterização acústica deste estado (Veiga, 2013). Por se tratar de um sistema com amplo vocabulário, as palavras modeladas são decompostas em suas respectivas transcrições fonéticas.

Cada fonema é representado por uma HMM que contém três estados emissores e uma topologia simples do tipo esquerda-direita que é comumente utilizada em sistemas ASR (Veiga, 2013) como ilustrado na Figura 2.2. Estados de entrada e saída, não emissores, são acrescentados para auxiliar na junção de modelos. Assim, o estado de saída de um fonema pode se juntar ao estado de entrada de outro fonema criando um HMM composto. Isto permite que os modelos se juntem formando palavras e estas unidas formando frases.

O objetivo da modelagem acústica é definir quais as sequências de estados, e com isto a sequência de fonemas, possuem uma melhor verossimilhança a um vetor de características acústicas que está sendo processado, definindo assim uma série de probabilidades de palavras e sequências de palavras para o segmento.

Os modelos acústicos geralmente utilizam o HMM para lidar com a variabilidade temporal da fala e os Modelos de Misturas de Gaussianas (GMM - Gaussian Mixture Model) na representação da relação entre os estados do HMM e a entrada acústica. Uma abordagem alternativa de representação da relação é o uso de uma rede neural feed-forward, a qual recebe vários quadros de coeficientes de entrada e produz probabilidade sobre os estados HMM como saída. Esses modelos baseados em HMM e redes neurais profundas (DNN - Deep Neural Networks), mostraram superar o GMM em diversos benchmarks para reconhecimento automático de fala [Hinton et al., 2012].

Apesar de todas a vantagens que os GMM's trouxeram para o reconhecimento de fala, eles apresentam uma grave deficiência: são estatisticamente ineficientes para modelar dados que se situam sobre ou próximo de um coletor não-linear no espaço de dados. Isso implica diretamente no entendimento da estrutura que a fala produz. Com isso, acredita-se que outros tipos de modelos, como o de redes neurais, podem possuir resultados melhores para a modelagem acústico se for possível explorar de forma mais efetiva informações em uma grande janela de quadros da estrada acústica. As redes neurais têm a capacidade de extração de informação de dados sobre, ou próximo de um, coletor não-linear muito melhor que o GMM.

A técnica para treinamento de modelos acústicos usando redes neurais faz uso dos novos algoritmos de aprendizagem de máquinas e a capacidade de hardware para a formação de redes neurais profundas que contém muitas camadas de unidades não-lineares escondidas e uma camada de saída grande. Esta camada de saída grande é necessária para acomodar o grande número de estados do HMM que surgem quando cada fonema é modelado a partir das ligações dos modelos HMM's de fonemas que surgem no decorrer do treinamento.

Modelo de Linguagem

O modelo de linguagem (LM - *Language Model*) não é necessário para o pleno funcionamento de sistemas ASR. Porém, utilizar somente informações acústicas não é suficiente para o desempenho satisfatório destes sistemas, uma vez que uma palavra poderia ser seguida por qualquer outra sem restrição (Silva, 2010). Assim, é consenso na área a

utilização de modelos de linguagem para sistemas que lidam com amplo vocabulário, trazendo vantagens para o processo, como a redução do espaço de busca, redução do tempo de reconhecimento e a resolução de ambiguidades acústicas (Pessoa et al., 1999; Silva, 2010).

O propósito do LM é prover a probabilidade de ocorrência de uma palavra $P(w_k)$ dada uma sequência de palavras que a antecedem, como é descrito na fórmula abaixo:

$$P(w_k) = P(w_k)P(w_k|w_{k-1})P(w_k|w_{k-1}w_{k-2})...P(w_k|w_{k-1}w_{k-2}...w_1) \tag{2.1}$$

Onde w_k, é uma palavra da sequência de palavras. Assim, a probabilidade de uma sequência de palavras w_k^1 ocorrer é dado por:

$$P(w_k^1) = P(w_1)\prod_{i=2}^{k} P(w_i|w_1...w_{i-1}) \tag{2.2}$$

Pode-se observar que o cálculo destas probabilidades para grandes vocabulários é algo inviável, pois a probabilidade de ocorrência será calculada em relação a todas as palavras da frase. Para resolver este problema é utilizado n-gramas, onde se assume a probabilidade de uma dada palavra somente até n-1 palavras que a antecedem. Assim, o cálculo de $P(W_k^1)$ é dado por:

$$P(W_k^1) = P(w_1)\prod_{i=2}^{k} P(w_i|w_{i-n+1}) \tag{2.3}$$

Os n-gramas são bastante eficientes em línguas onde a ordem das palavras é importante, pois o modelo de linguagem abrange características de sintaxe e semântica e evita a necessidade da criação de regras e de uma gramática formal. Em sistemas ASR é comumente utilizado trigramas (n = 3) e 4-gram (n = 4), pois a maioria das palavras possuem forte dependências das duas que a antecedem.

Para avaliação de um modelo de linguagem, utiliza-se a medida de perplexidade. A perplexidade representa o quão indeciso o modelo pode ficar dada uma sequência textual. Quanto menor o valor da perplexidade, melhor é o modelo (Young et al., 2006). Esta medida é obtida utilizando uma base de teste contendo frases, onde cada frase é avaliada segundo o modelo de linguagem construído, com isso um bom modelo de linguagem

deve apresentar uma perplexidade alta para uma frase sintaticamente incorreta e uma perplexidade baixa para uma frase sintaticamente correta.

Dicionário fonético

O Dicionário Fonético ou Léxico é composto basicamente por um sequência de palavras e suas respectivas transcrições fonéticas. Esta técnica também é conhecida como conversão grafema fonema, ou seja, conversão de símbolos gráficos utilizados na construção de palavras da língua em unidades sonoras usadas para formar e distinguir palavras.

Este componente é um importante pré-requisito para a construção de sistemas ASR. Porém, a conversão de uma sequência de caracteres em sequência de fones não é uma tarefa trivial. Sabe-se que a transcrição fonética gerada automaticamente apresenta algumas limitações. Uma dessas fraquezas é a dificuldade em tratar as combinações formadas pelo final de uma palavra com início de uma outra. Por exemplo, a palavra "dois" termina com um "s" e soa como "ss" sempre, exceto quando a palavra falada após o "dois" começar com vogal. Nesse caso, o último fonema de dois se transforma em "z". Há diversos trabalhos que lidam com esses e outros problemas relacionados à transcrição fonética automática.

Decodificador

O decodificador é o componente dos sistemas ASR responsável por manipular todos os outros componentes descritos acima e gerar a melhorar sequência textual aproximada para um dado segmento de áudio de entrada.

A tarefa do decodificador é obter a melhor sequência de palavras que descreve uma dada sequência de características acústicas (gerada pelo *front-end* a partir de um dado áudio). O processo de decodificação é construído por um rede de palavras geradas pela modelagem do sistema. Esta rede é definida por um conjunto de nós (palavras) conectados por arcos, onde cada arco possui uma probabilidade de ocorrência (transição) (Silva, 2010). As palavras são representadas por sequências de fonemas oriundos do modelos acústicos e a escolha das palavras obedece ao modelo de linguagem que define um universo de hipóteses para as sequências de palavras (Veiga, 2013). Após encontrar o

melhor caminho na rede(aquele cuja probabilidade é mais alta), o decodificador transcreve a sequência de fonemas encontrados em suas respectivas representações grafemáticas no dicionário fonético e por fim gera a sequência textual, em linguagem natural, e finaliza o sistema.

2.2 Treinamento de modelos acústicos

O treinamento de modelos acústicos tem como objetivo a geração de modelos para sistemas ASR influenciando em um acurácia satisfatória destes. Para o treinamento de modelos acústicos HMM é necessário o fornecimento das seguintes informações: conjunto de arquivos de áudio bem segmentados; conjunto de arquivos de texto com as respectivas transcrições originais dos arquivos de áudio; dicionário de abrangência do modelo e respectivas transcrições fonéticas. A partir disto, os passos para o treinamento de modelos acústicos são:

1. Extração de parâmetros dos arquivos separados para treino: nesta etapa são geradas características acústicas importantes para o treinamento

2. Inicialização dos modelos de fonemas: neste ponto, utiliza-se da técnica *Flat Start* para segmentação automática da base. Esta técnica é dividida em três etapas: a primeira é desconsiderar a pausa existente entre as palavras, gerando uma primeira estimativa dos modelos de fonemas. O segundo é a criação de um modelo de pausa a partir do modelo de silêncio de início e fim de frase (justificando um base de dados bem segmentada, ou seja, respeitando início e fim de frase), após isto reestimam-se os modelos. O terceiro e último é o realinhamento da base de treino em função dos modelos e estes são reestimados novamente (Tevah, 2006; Young et al., 2006).

3. Conversão do modelo de fonemas para trifones: os arquivos de texto são novamente transcritos agora com trifones, levando-se ou não em conta, os trifones entre-palavras. O fones centrais dos trifones referenciam os fonemas treinados no passo anterior (Silva, 2010).

4. Compartilhamento dos estados a partir de árvores de decisão: a lista de todos os

trifones necessários ao modelo acústico é criada e no final é gerado um arquivo mapeando modelos que compartilham distribuições (Tevah, 2006; Young et al., 2006).

5. Aumento do número de Gaussianas dos modelos: neste passo são retreinados os modelos, dos passos anteriores que possuem apenas uma Gaussiana, para cada nova Gaussiana adicionada (Silva, 2010).

2.3 Treinamento de modelos de linguagem

O treinamento de modelos de linguagem tem como objetivo a geração de modelos para sistemas ASR que auxiliarão no reconhecimento de fala e na construção gramatical dos resultados. Para o treinamento do modelo de linguagem são requeridos o dicionário fonético e os arquivos de texto de onde serão extraídas as frequências de palavras e o relacionamento entre elas.

Algumas alterações são necessárias nos arquivos de texto para que o treinamento seja satisfatório. É de fundamental importância que os arquivos de texto tenha marcadores de início e final de sentença (s, /s) e que as frases estejam normalizadas(números e siglas) e livres de pontuações(. , ! ? ... etc) (Young et al., 2006).

A seguir serão descritos os passos para a criação de modelos de linguagem do tipo unigrama, bigrama e trigrama.

1. Geração da frequência das palavras existentes nos arquivos de texto e do relacionamento entre cada palavra com até duas anteriores, incluindo início e final de frase (Tevah, 2006; Young et al., 2006).

2. Identificação de palavras fora do vocabulário (OOV - *Out of Vocabulary*) que são excluídas do treinamento.

3. Treinamento dos n-gramas: Computação dos modelos unigrama, bigrama e trigrama incluindo probabilidades de sequências de palavras e fatores de escalonamento (Tevah, 2006; Young et al., 2006).

4. Medição da perplexidade dos modelos n-gramas gerados.

2.4 Base de dados

Para obter sistemas ASR que sejam robustos e com acurácia satisfatória é necessário uma base de dados bem estruturada e de grande porte para o treinamento de modelos acústicos e de linguagem.

O termo *corpus* (plural *corpora*) de voz será utilizado neste trabalho para representar a base de dados de áudio com suas respectivas transcrições textuais, utilizadas no treinamento de modelos acústicos e o *corpus* de texto para representar a base de dados de sentenças (frases) estruturadas, utilizadas para treinar o modelo de linguagem.

Uma grande dificuldade no treinamento de modelos acústicos e de linguagem para o PB é a obtenção destes *corpora* de grande porte e gratuitos (Silva et al., 2005). Isto é uma carência na área que difere de outras línguas como o inglês. Além disto, é importante que estes *corpora* sejam especializados, ou seja, que áudios e textos contendo informações de diferentes áreas de conhecimento sejam utilizados no treinamento. Isto para que o sistema não seja "viciado" em termos de uma única área (a não ser que este seja o foco) ou de nenhuma.

Além de possuir *corpora* que sejam suficientemente grande, outras questões são levantadas sobre este fator de vital importância na construção de sistemas ASR. É necessário que os textos estejam normalizados, ou seja, dígitos e siglas sejam traduzidos para suas respectivas representações textuais, também é de vital importância realizar uma varredura palavra por palavra para validação ortográfica (são utilizados corretores ortográficos para esta tarefa) (Tevah, 2006) e que as frases sejam foneticamente balanceadas(distribuição fonética similar à encontrada na fala) (Ynoguti, 1999). Ainda, para a base de áudio é necessário que o começo e fim de segmentos sejam respeitados com silêncio e que os arquivos de áudio tenham uma mesma configuração (amostragem, quantidade de canais e *codec*).

2.5 Trabalhos relacionados

No contexto deste trabalho, não foram encontradas conclusões na literatura de quais tarefas do treinamento de modelos acústicos e de linguagem possuem maior ganho e são

críticas nestes processos. Porém, existem diversos trabalhos que propõem a utilização de técnicas específicas ou alterações das abordagens tradicionais de treinamento que influenciam na acurácia de modelos.

2.5.1 Modelos acústicos

O modelo acústico é, com certeza, o principal fator de relevância nos resultados de sistemas ASR (Oliveira et al., 2011). Bons modelos acústicos, associados a demais fatores, geram resultados satisfatórios para estes sistemas.

Visando esta melhora, Hinton et al. (2012) apresenta a revisão da técnica de utilização de DNN (*Deep Neural Networks*) no treinamento de modelos acústicos, para substituir a abordagem tradicional de GMM (*Gaussian mixture model*), na representação da relação entre os estados do HMM e a entrada acústica. Esta técnica, basicamente, utiliza-se dos novos algoritmos de aprendizagem de máquinas e a capacidade de hardware para a formação de redes neurais profundas que contém muitas camadas de unidades não-lineares escondidas e uma camada de saída grande. Esta camada de saída grande é necessária para acomodar o grande número de estados do HMM que surgem quando cada fonema é modelado e a partir das ligações dos modelos HMM's de fonemas que surgem no decorrer do treinamento. O ajuste dos DNN's no treinamento é realizado em duas etapas. Na primeira etapa são inicializados os detectores de características, uma camada de cada vez, criando uma pilha de modelos, cada um dos quais tem uma camada de variáveis latentes. Na segunda fase, cada modelo da pilha é usado para inicializar uma camada de unidades ocultas em um DNN e então a rede é aperfeiçoada para prever os estados alvos. Estes alvos são obtidos utilizando um sistema base HMM-GMM para o alinhamento forçado. Esta técnica é compartilhada por grupos de pesquisa da Universidade de Toronto, Microsoft Research , Google and IBM Research.

Em (Thomas et al., 2013) também é apresentado técnicas de utilização de redes neurais para melhorar a acurácia de modelos acústicos e demonstrado que existe uma melhora significativa nos modelos. Além da utilização de DNN para representação da relação entre os estados HMM e a entrada acústica, também é proposto no trabalho a criação de um *front-end* para a extração de características baseado em redes neurais.

Em (Ma et al., 2006) é avaliado o treino de modelos com quantidades variáveis de dados, onde são testados vários modelos de linguagem (genéricos e contextual) e métodos de treino de modelos acústicos (ML, MMI e treino com adaptação de locutor).

2.5.2 Modelos de linguagem

Com o intuito de investigação das técnicas para a modelagem de linguagem, Silva et al. (2004) descrevem as diversas propostas de construção e melhorias de modelos, dificuldades e resultados comparativos. Nesse trabalho é apresentado ainda umas das principais dificuldades na geração das probabilidades de palavras, que provém da esparsidade dos dados do treinamento. Isto é chamado de problema de frequência zero, que acontece quando palavras com probabilidade zero nunca serão reconhecidos mesmo sendo acústicamente plausíveis. Para resolver este problema é proposto técnicas de suavização buscando assegurar que as palavras possuam probabilidades positivas. Além disso, é apresentado alguns métodos para balanceamento da distribuição do contexto total. Entre eles, o desconto linear e absoluto, onde a estimativa de máxima verossimilhança dos contextos são descontadas proporcionalmente às probabilidades, e suas variações. Também apresenta o modelo aditivo interpolado onde, para cada contexto de palavras do treinamento, usa-se uma constante aditiva para suavizar a distribuição. No trabalho é concluído que, na comparação entre as técnicas, o modelo aditivo interpolado apresenta os melhores resultados com custo computacional relativamente baixo.

2.5.3 Dados para treino

Um ponto também importante e de significativa influência na acurácia de modelos é o conteúdo utilizado no treinamento. A dificuldade de possuir o *corpus* de voz e texto adequados é sem dúvida um dos principais problemas encontrados pelos pesquisadores da área.

Sistema típicos de reconhecimento de fala são treinados usando centenas de horas de dados de treinamento acústico cuidadosamente transcritos, como é dito em (Evermann et al., 2005). Esse trabalho descreve as dificuldades no reconhecimento automático de fala para sistemas CTS (*Conversation Telephone Speech*) e como o aumento da base de

dados em milhares de horas melhora a acurácia dos modelos acústicos e de linguagem. Mais especificamente o trabalho compara dois sistemas de transcrição treinados com as mesmas tecnicas e quantidades de dados diferentes, um com 360 e o outro com 2000 horas de conversas telefônicas e descreve quais os métodos de limpeza nos dados de áudio e de transcrições associados. Foram aplicados regras para normalizar o texto, corrigir ortografia e descartar palavras que foram ditas somente uma vez. Além disso foi mapeado, na base de áudio, blocos de silêncio e falas masculinas e femininas. O trabalho conclui apresentando as melhoras que foram adquiridas no incremento da base e os valores de ganho na acurácia e taxa de erro de palavras (WER).

Para o PB existe uma dificuldade grande em possuir uma base extensa e gratuita. Visando contornar este problema, Wessel e Ney (2005) expõem um estudo da influência da quantidade de dados transcritos, inicialmente disponíveis, no desempenho dos modelos acústicos. Nesse trabalho é proposto um método diferente para o treinamento de modelos com poucos dados iniciais. Visto que, existe muita quantidade de dados acústicos e poucos desses estão associados à transcrições, o trabalho utiliza-se de poucos dados transcritos manualmente para treinar um sistema de base que será usado para reconhecer outros dados acústicos e assim utiliza-los também no treinamento. Para validação da proposta é feita uma marcação de confiança nas palavras transcritas, ou seja, cada palavra transcrita pelo sistema base é associada a uma medida de confiança, identificando que somente palavras acima de um limiar serão utilizadas no treinamento. Nessa abordagem, primeiramente, é estimado os parâmetros de treinamento de um modelo acústico com pequenas quantidades de fala transcritas manualmente. Este modelo acústico é então usado para reconhecer *corpora* de voz grandes que não possuem transcrições associadas e gerar um gráfico de palavras, que é computado a fim de gerar as medidas de confiança. Assim um novo *corpus* de voz é criado unindo aquele utilizado no primeiro treinamento e o novo gerado. Com isso, o mesmo procedimento padrão de treinamento é executado com o *corpus* combinado às medidas de confiança geradas. É ressaltado ainda que este processo gera melhoras nas medidas de confiança conforme novos modelos vão surgindo. Ainda no trabalho é apresentado os ganhos obtidos na acurácia dos modelos e na taxa de erro de palavras (WER) a partir dos experimentos realizados.

Existem ainda outros trabalhos que abordam o problema de treinar modelos com poucos recursos. Em (Riccardi e Hakkani-Tür, 2003) é apresentada também uma técnica para transcrever e gerar uma medida de confiança para as transcrições. Os segmentos que não possuem transcrições com boa confiança são transcritos manualmente construindo assim uma base de dados com qualidade melhor. Nesse trabalho é descrito um método para combinar aprendizagem ativa e não supervisionada para um sistema de reconhecimento de fala. O objetivo é minimizar a supervisão humana para o treinamento de modelos acústicos e de linguagem e maximizar o desempenho com os dados de áudio com transcrições e sem transcrições associadas. A aprendizagem ativa visa diminuir o número de dados de áudio a serem transcritos manualmente processando-os e selecionando os mais precisos com relação a uma função de custo. Para a aprendizagem não supervisionada foi utilizada as transcrições automáticas e uma medida de confiança por palavra. O trabalho é concluído apresentando os ganhos e benefícios da combinação de aprendizagem ativa e não supervisionada.

Há também trabalhos para o aproveitamento de mídias legendadas como em (Chan e Woodland, 2004), onde é aproveitada a disponibilidade de legendas em noticiários para treinar modelos acústicos e gerar frases foneticamente balanceadas para o modelo de linguagem. Nesse trabalho é apresentado os problemas em se utilizar legendas como base para treino. O principal deles é que, as transcrições geralmente são oriundas de *closed caption* e legendas de comerciais, que são construídos manualmente e são parcialmente corretos. Embora estes transcritos contenham uma série de erros e não possam ser usados diretamente como dados para treino, é proposto no trabalho o usos destes dados como fonte de supervisão para o treino de modelos acústicos, técnica conhecida como treinamento levemente supervisionado. Na abordagem proposta do treinamento levemente supervisionado, foi utilizado modelos de linguagem tendenciosos para as transcrições de *closed caption* que foram usados no processamento da base de áudio. Com isso, todas as transcrições adquiridas foram utilizadas para treinamento com a técnica de máxima verossimilhança ou para fornecer as transcrições corretas no treinamento discriminativo. Os resultados apresentados mostram que é possível a diminuição da taxa de erro de palavras utilizando esta abordagem e que a construção de uma base de treinamento com este tipo

de mídia é viável para um sistema de reconhecimento dentro desse contexto.

Outro trabalho relacionado a mídias legendas é apresentado em (Panayotov et al., 2015), onde é aproveitado o conteúdo de *audio books* para alimentar a base de dados de treino. Nesse trabalho é proposto a criação de uma base de treino para modelos de sistemas de reconhecimento de fala, utilizando conteúdo de *audio books* em inglês disponibilizados gratuitamente pelo projeto LibriVox[1]. A grande questão de se utilizar *audio books* para o treinamento de modelos, é o alinhamento correto do áudio com o texto escrito. Os processos de treinamento, geralmente, requerem que os dados venham segmentados até algumas dezenas de segundos e contenham o texto relacionado ao segmento. A abordagem proposta no trabalho é composta de dois estágios. No primeiro estágio é utilizado um algoritmo de alinhamento para encontrar a melhor região entre o áudio reconhecido e o texto do capítulo. A partir disso é tomada a maior região de similaridade, que geralmente representa um capítulo inteiro, e descartado o resto. Nessa região de similaridade é destacadas palavras que, a partir do reconhecimento, tenham uma alta confiança. Depois disso o áudio é dividido em segmentos de algumas dezenas de segundos utilizando um algoritmo de programação dinâmica. Esta divisão é feita a partir do reconhecimento de regiões de silêncio. Assim é gerado um texto candidato para um segmento de áudio. Na segundo etapa é filtrado os segmentos onde o texto candidato tem uma alta probabilidade de ser impreciso. Também, é criado um gráfico de decodificação para cada segmento. Nesta etapa é rejeitado qualquer segmento que tenha desvio da transcrição original. Isto é feito utilizando técnicas para comparar palavras com blocos de áudio de fala e identificar variações bruscas dos fones de cada palavra e do sinal processado. Além disso foi filtrado segmentos com áudios barulhentos (ruído). O trabalho conclui demonstrado o resultado superior do sistema construído a partir de *audio books* em comparação a outro sistema treinado com bases de dados tradicionais, transcritos manualmente, porém de menor tamanho.

Existem ainda, bases gratuitas que foram construídas por trabalhos acadêmicos, como no desenvolvido pelo Laboratório de Processamento de Sinais (LaPS[1]) da Universidade Federal do Pará, onde foi construído um *corpus* de voz de aproximadamente 11

[1] `https://librivox.org`

[1] `http://www.laps.ufpa.br/falabrasil/descricao.php`

horas de áudio e um *corpus* de texto com 120 mil frases que são disponibilizados gratuitamente. Outra base disponibilizada gratuitamente é o CETENFolha, que consiste em textos extraídos do jornal Folha de São Paulo e disponibilizados gratuitamente em formato adequado para processamento de texto. A base é mantida pelo projeto Linguateca[2] e contém quase 2 milhões de frases em formato ideal para o processamento de modelos de linguagem.

2.5.4 Conversores fonéticos

Existem ainda trabalhos relacionados com a criação do dicionário fonético a as dificuldades da tarefa de conversão de sequência de caracteres em sequência de fones como em (Siravenha et al., 2008), onde são apresentadas as duas técnicas de destaques na literatura para a conversão grafema fonema: *data-driven* e baseadas em regras. O trabalho apresenta primeiramente a importância da elaboração de um dicionário fonético de grande vocabulário na construção de um sistema de reconhecimento automático de voz e como este componente, quando bem construído, impacta positivamente na acurácia dos modelos. Com isso, é proposto a contribuição de um algoritmo baseado em regras para a conversão grafema fonema. Os conversores baseados em regras apresentam a vantagem de não fazer uso de alinhamento lexical, visto que não há necessidade de treinamento para gerar as próprias regras. A proposta do trabalho é fornecer ao algoritmo, regras fonológicas pré-estabelecidas de acordo com a língua no qual o sistema é baseado. Por fim é apresentado os desafios e resultados da abordagem proposta, bem como os ganhos atingidos em comparação com as abordagens tradicionais. Os melhores resultados foram obtidos em sistemas que utilizaram um dicionário baseado em regras, apresentando a menor taxa de WER (*Word Error Rate*).

2.5.5 Conclusão

Estes esforços da comunidade se relacionam com o presente trabalho na tentativa de encontrar novas soluções, técnicas e dados para o treinamentos de modelos acústicos e de linguagem que possibilitam melhorias na acurácia de sistemas ASR. O objetivo é utilizar os

[2] `http://www.linguateca.pt/ACDC/`

trabalhos apresentados para tomar uma linha de base nos experimentos e na investigação dos processos.

3 Experimentos

O conjunto de experimentos realizados neste trabalho tem como objetivo indicar quais etapas na construção de modelos para um sistema de reconhecimento automático de fala possuem maior importância e relevância na acurácia final dos modelos. Estes sistemas de reconhecimento de fala utilizam dois modelos: modelo de linguagem e modelo acústico. Neste capitulo serão descritos quais o modelos gerados e a combinação de modelos avaliados pelo trabalho. Ainda, será descrito a base de treinamento e de avaliação utilizada nos experimentos.

3.1 Base de Treinamento

A base de dados utilizada no treinamento de modelos é um dos principais elementos de influência na acurácia dos sistemas de reconhecimento de fala. O ideal para esta base de dados é que seja o mais abrangente possível com relação a fala humana (sotaque, entonação, gênero do falante etc), aos diversos contextos de assuntos existentes (informática, saúde, política etc) e às características de criação do conteúdo (videoaula, filme, novela, jornal etc). Contudo, ter uma base de dados com essas características e que seja gratuita é um problema da área. Assim, os experimentos foram realizados utilizando uma base de dados especializada.

O *corpus* de voz foi construído, principalmente, a partir de videoulas legendadas, disponibilizadas de forma gratuita no repositório acadêmico do Coursera[1], em português do Brasil. O tempo total do *corpus* construído foi de aproximadamente 55 horas de videoaulas. Todas as videoaulas foram gravadas em ambiente apropriado e possuem uma boa qualidade de áudio. O assuntos abordados nas videoaulas são: Informática, Ensino, Gestão, Empreendedorismo e Biologia. Também foram utilizados *corpora* de voz gratuitos, disponibilizadas pelo projeto VoxForge[2] e pelo Laboratório de Processamento de Sinais

[1] `https://pt.coursera.org/`
[2] `http://www.voxforge.org/home`

(LaPS) da UFPA[3], que juntas somam 2 horas de áudio. Com isso, o *corpus* de voz conta com 57 horas de áudio.

O *corpus* de texto foi construído a partir do texto das legendas do *corpus* de voz do Coursera e de bases gratuitas como CETEN, OGI e LapsFolha disponibilizadas pelo Laboratório de Processamento de Sinais (LaPS) da UFPA e dos artigos da Wikipédia[4] em português do Brasil. O *corpus* completo conta com mais de 13 milhões de frases.

3.2 Avaliação de Sistemas de Reconhecimento Automático de Fala

Nesta seção será descrita a base utilizada na avaliação dos experimentos e as métricas utilizadas.

3.2.1 Base de Avaliação

Para avaliação de sistemas de reconhecimento automático de fala, utiliza-se um *corpus* de voz, que chamaremos de referência, cuja transcrição é exata com o áudio e que não tenha sido utilizado no treinamento dos modelos avaliados. O *corpus* utilizado neste trabalho foi construído manualmente a partir de videoaulas do repositório acadêmico videoaula@RNP[1] da Rede Nacional de Ensino e Pesquisa. Este *corpus* totaliza cerca de 2 horas de áudio e 581 frases que foram transcritas manualmente. A base de avaliação foi construída desta forma pois apresenta o mesmo contexto de criação (videoaulas) do *corpus* utilizado no treinamento do sistema avaliado. O tempo gasto na geração desta base foi necessário pois não foi encontrada uma base de avaliação disponível na Web para o contexto de criação do modelo. Foram necessárias duas semanas para criação da base manual para avaliação dos modelos.

[3] `http://www.laps.ufpa.br/falabrasil/`
[4] `https://www.wikipedia.org/`
[1] `http://www.videoaula.rnp.br/portal/home`

3.2.2 Métricas

Os sistemas de reconhecimento de fala são comumente avaliados segundo a taxa de erro de palavras (WER - Word Error Rate). Esta métrica representa o quanto a transcrição gerada pelo sistema de reconhecimento de fala (*hypotheses*) difere da transcrição original (*reference*).

A WER é calculada a partir da definição de quais palavras da hipótese estão corretas, quais foram inseridas incorretamente, quais foram excluídas e quais foram substituídas em comparação com a referência. A WER é calculada pela seguinte fórmula:

$$WER = 100 * \frac{I + S + D}{total\ de\ palavras\ em\ reference} \tag{3.1}$$

Onde I é o número de palavras que foram inseridas incorretamente, S é o número de palavras que foram substituídas incorretamente e D é o número de palavras que foram deletadas incorretamente.

Os modelos de linguagem, por serem bastante utilizados em outras aplicações, podem ser avaliados separadamente a partir da métrica de perplexidade. A perplexidade do modelo de linguagem em um conjunto de avaliação pode ser definido como o inverso da probabilidade de ocorrência do conjunto de avaliação normalizado pelo número de palavras. Para um conjunto de avaliação W formado por palavras, tal que $W = w_1 w_2 ... w_N$, a perplexidade pode ser definida como:

$$Perplexidade(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}} \tag{3.2}$$

O formato do calculo da probabilidade é dependente da ordem do modelo de linguagem que está sendo avaliado. Assim, a fórmula acima é para modelos de linguagem unigram. Para o modelo bigrama a fórmula é a seguinte:

$$Perplexidade(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}} \tag{3.3}$$

3.3 Modelo de linguagem

O modelo de linguagem não é um elemento exclusivo de sistemas de reconhecimento de fala. Ele é comumente utilizado em diversas aplicações que envolvam PLN.

Para o modelo de linguagem existem diversos experimentos possíveis para alterar a construção do modelo de linguagem e avaliar seus resultados. Serão avaliadas 4 variáveis para o treinamento de modelos de linguagem: ordem, tamanho, processamento de texto e método de desconto.

A ordem está relacionado a dependência de cada palavra dado as $n-1$ palavras que a antecedem. Isto quer dizer que, para um modelo de ordem 3 (trigram), a probabilidade de ocorrência de uma palavra tem relação com as duas que a antecedem. Para o modelo de ordem 4 (4-gram), a probabilidade de ocorrência de uma palavra tem relação com as três que a antecedem. Nos experimentos realizados, foram utilizados modelos trigram e 4-gram interpolados. Interpolados quer dizer que foram gerados os modelos até 3-gram e até 4-gram. Assim para o modelo 4-gram por exemplo, foram gerados os modelos 1-gram, 2-gram, 3-gram e 4-gram.

O tamanho está relacionado à quantidade de texto que foi utilizado para o treinamento do modelo. Este é um importante quesito para modelos de linguagem pois, dependendo da aplicação, é necessário que a abrangência de palavras e de contextos de utilização das palavras sejam grandes. Nos experimentos realizados, foi avaliado a inserção de uma grandes fontes de texto, o Wikipédia.

O processamento de texto foi realizado para padronizar os textos da base de treinamento e ajustar o modelo. O principal ajuste realizado foi a expansão de padrões em sequência de palavras. Dentre os padrões expandidos estão: números cardinais (ex: 12, 0.3); números ordinais (ex: Bill Gates III, 1º lugar); números romanos (ex: século XIX); datas e horas (ex: 11:45, 28/01/2017) ; abreviações (ex: Av, LTDA etc); sequências de letras (ex: DVD, IBM, PC); porcentagem (ex: 75%, 100%); valores monetários (ex: R$23.42, $3.45). Isto foi realizado para adequar o texto à forma humana como é falado. Nos experimentos realizados foi avaliado o impacto da normalização do texto na avaliação do modelo de linguagem.

O método de desconto está relacionado a suavização das probabilidades do modelo

de linguagem. Os métodos de descontos são utilizados de forma recorrente na criação de modelos de linguagem para grandes vocabulários. O objetivo dos métodos de descontos é impedir que algumas palavras, que possuem a frequência baixa na base de treinamento, tenham probabilidade zero. Assim, uma parte da probabilidade da massa de palavras mais frequentes é realocada para estas palavras com probabilidade zero. Desta forma estas palavras deixam de ser invisíveis no modelo de linguagem. Nos experimentos foi avaliado o uso dos métodos Good-Turing (Katz, 1987), Chen and Goodman (Chen and Goodman, 1996) e Kneser-Ney (Ney and Essen, 1991).

Foram realizados 5 experimentos, a partir das variáveis descritas acima, para o modelo de linguagem, e estão descritos abaixo.

- **LM1:** Neste experimento foram gerados dois modelos de linguagem (3-gram e 4-gram), que foram treinados utilizando 80% do texto de legenda das videoaulas da base de áudio, sem nenhum tipo de processamento do texto ou normalização e método de desconto Good Turing. A base de avaliação (teste) do modelo gerado foi criada utilizando o restante do texto de legendas das videoaulas da base de áudio (20%).

- **LM2:** Neste experimento foram gerados dois modelos de linguagem (3-gram e 4-gram interpolados), que foram treinados utilizando 60% do texto de legenda das videoaulas da base de áudio junto com os dataset CETEN, OGI, LapsFolha e Wikipédia, sem nenhum tipo de processamento do texto ou normalização e método de desconto Good Turing. A base de avaliação (teste) do modelo gerado foi criada utilizando o restante do texto de legendas das videoaulas da base de áudio (40%).

- **LM3:** Neste experimento foram gerados dois modelos de linguagem (3-gram e 4-gram interpolados), que foram treinados utilizando 60% do texto de legenda das videoaulas da base de áudio junto com os dataset CETEN, OGI, LapsFolha e Wikipédia, com texto processado e normalizado e método de desconto Good Turing. A base de avaliação (teste) do modelo gerado foi criada utilizando o restante do texto de legendas das videoaulas da base de áudio (40%).

- **LM4:** Neste experimento foram gerados dois modelos de linguagem (3-gram e 4-

gram interpolados), que foram treinados utilizando 60% do texto de legenda das videoaulas da base de áudio junto com os dataset CETEN, OGI, LapsFolha e Wikipédia, com texto processado e normalizado e método de desconto Chen and Goodman. A base de avaliação (teste) do modelo gerado foi criada utilizando o restante do texto de legendas das videoaulas da base de áudio (40%).

- **LM5:** Neste experimento foram gerados dois modelos de linguagem (3-gram e 4-gram interpolados), que foram treinados utilizando 60% do texto de legenda das videoaulas da base de áudio junto com os dataset CETEN, OGI, LapsFolha e Wikipédia, com texto processado e normalizado e método de desconto Kneser-Ney. A base de avaliação (teste) do modelo gerado foi criada utilizando o restante do texto de legendas das videoaulas da base de áudio (40%).

Os experimentos com o modelo de linguagem foram reproduzidos utilizando o *toolkit* SRILM (Stolcke et al., 2002), que possui ferramentas para treinamento de modelos e avaliação de perplexidade.

3.4 Modelo acústico

O modelo acústico é o principal elemento de sistemas de reconhecimento de fala. O treinamento deste modelo utiliza o *corpus* de voz para geração de probabilidades a partir das características acústicas observadas.

Para os experimentos com o modelo acústico foi avaliada a estrutura da modelagem, o alinhamento do *corpus* e as configurações do áudio utilizado no treinamento.

A avaliação do alinhamento do *corpus* foi necessária pois foram encontrados alguns erros de alinhamento do áudio com a legenda nas videoaulas extraídas do Coursera. Para quebrar o áudio em segmentos foi utilizada a marcação de tempo de legenda que cada videoaula possuía (arquivo .vtt), porém as marcações geralmente estavam imprecisas e mal alinhadas com o áudio. Para resolver este problema foi desenvolvido um método de alinhamento dos áudios (Ferreira e Souza, 2017) utilizando a ferramenta WebRTC Voice Activity Detector[1] (VAD) como apoio. O alinhamento utiliza a ferramenta para detectar

[1] `https://github.com/wiseman/py-webrtcvad`

os intervalos do áudio onde existe atividade de fala. A partir da detecção desses intervalos os segmentos foram gerados concatenando os trechos de legenda disponibilizados, limitados por um determinado limiar α (1 segundo), que representa a diferença que poderia haver entre o tempo final de um intervalo de fala e o intervalo final da marcação de legenda. O Algoritmo 1 apresenta o alinhador.

Algoritmo 1: Alinhar áudio a texto da legenda

```
Saída: Áudios segmentados
1  begin
2      S ←tempos originais das legendas;
3      V ←tempos com algoritmo VAD;
4      α ← 1;
5      for cada V_i ∈ V do
6          t_beg ← V_i0;
7          while V_i1 ∉ [S_j1 − α, S_j1 + α] e V_i1 ⩽ S_j1 + α do
8              i++;
9              if V_i1 ⩾ S_j1 + α then
10                 j++;
11             end
12         end
13         t_end ← V_i1;
14         segmenta(t_beg, t_end);
15     end
16 end
```

Nesse pseudocódigo, as variáveis S e V são estruturas que armazenam os tempos gerados pelo arquivo de legenda e o algoritmo VAD. Ambos contém, para cada linha, um tempo de início e de fim do segmento. Desta forma, por exemplo, V_{i0} e V_{i1} representam respectivamente o tempo inicial e final do segmento V_i tal que i varia de zero ao número de segmentos existentes. O objetivo deste alinhador é encontrar valores iniciais (t_beg) e finais (t_end) que indique um segmento melhor alinhado para o áudio, utilizando como auxílio o tempo da legendas e o algoritmo de segmentação por identificação de atividade de fala.

Assim, foram gerados 4 modelos acústicos, que são descritos abaixo.

- **AM1:** Neste experimento foi gerado um modelo acústico treinado com características originais dos áudios do *corpus* de voz: *codec* WAV, taxa de amostragem de 16000 Hz, canais do tipo MONO. O *corpus* de voz não passou por uma etapa de

processamento da segmentação, ou seja, foi utilizado sem qualquer processamento. A estrutura da modelagem é GMM (Gaussian mixture model).

- **AM2:** Neste experimento foi gerado um modelo acústico com as seguintes características de áudio: *codec* WAV, taxa de amostragem de 8000 Hz, canais do tipo MONO. O *corpus* de voz não passou por uma etapa de processamento da segmentação, ou seja, foi utilizado sem qualquer processamento. A estrutura da modelagem é DNN (Deep Neural Network).

- **AM3:** Neste experimento foi gerado um modelo acústico com as seguintes características de áudio: *codec* WAV, taxa de amostragem de 8000 Hz, canais do tipo MONO. A segmentação do áudio é alinhada e a estrutura da modelagem é DNN (Deep Neural Network).

- **AM4:** Neste experimento foi gerado um modelo acústico com as seguintes características de áudio: *codec* WAV, taxa de amostragem de 8000 Hz, canais do tipo MONO e reverberações e distorções de tempo e volume. A segmentação do áudio é alinhada e a estrutura da modelagem é DNN (Deep Neural Network).

Os experimentos com o modelo acústico foram reproduzidos pelo *toolkit* Kaldi (Povey et al., 2011), que foi responsável pelo treinamento dos modelos. Ainda, foi utilizado o FFMPEG (FFmpeg Developers, 2017) para alteração nas configurações de áudio do *corpus* de voz.

4 Resultados

A criação dos modelos é a tarefa com o maior gasto de tempo em relação a todas as outras, por isso, é essencial que o maquinário disponibilizado para este fim seja adequado. Os experimentos para treinamento de modelos acústicos e de linguagem foram executados em uma máquina Linux Ubuntu 14.04, processador Intel(R) Core(R) i7, 16GB de RAM e GPU Nvidia GTX-970. Para gerar os resultados de WER e perplexidade era necessário um sistema que apenas mantivesse o decodificador e os modelos gerados. Assim, para geração dos resultados foi utilizado um máquina Linux Ubuntu 14.04, processador Intel(R) Xeon(R) CPU E5-2650 e 16GB de RAM.

Neste capitulo serão apresentados os resultados e discussões para os experimentos com modelo de linguagem e da combinação de modelos acústicos e modelos de linguagem.

4.1 Tempo das tarefas

Como objetivo deste trabalho é indicar quais etapas (ou tarefas) do processo de treinamento de modelos acústicos e de linguagem requerem mais esforço e influenciam no resultado final, a indicação de tempo gasto em cada etapa é um importante quesito para este objetivo. A Tabela 4.1 apresenta o tempo de preparação e o tempo de execução que foram gastos nas principais etapas. O tempo de preparação está relacionado ao esforço humano gasto na tarefa e o tempo de execução está relacionado ao esforço computacional.

Tabela 4.1: Tempo das tarefas

	Tempo	
Tarefa	**Preparação**	**Execução**
Criação do corpus de voz e texto	4 dias	-
Segmentação do corpus de voz	1 dia	10 minutos
Alinhamento do corpus de voz	7 dias	10 minutos
Normalização do corpus de texto	7 dias	20 minutos
Configurações de áudio do corpus de voz	-	10 minutos
Treinamento de Modelo Acústico GMM	1 dia	2 dias
Treinamento de Modelo Acústico DNN	1 dia	5 dias
Treinamento de Modelo de Linguagem	-	10 minutos
Avaliação dos modelos	-	3 horas

Algumas tarefas estão ou sem tempo de preparação ou sem tempo de execução pois, para estas, não foi demandado tempo. O tempo de preparação das tarefas é maior, porém tarefas como criação da base, preparação do método de alinhamento e do normalizador de texto ocorreram uma única vez. Já o tempo de execução das tarefas foi presente em cada experimento executado, por exemplo, foram gerados 3 modelos acústicos usando a técnica DNN e cada um demandou 5 dias para execução.

4.2 Resultados do modelo de linguagem

O modelo de linguagem interfere no sistema de reconhecimento de fala no momento em que as palavras são agrupadas formando frases. Caso um sistema de reconhecimento de fala não possua um modelo de linguagem adequado para a aplicação, o modelo acústico não será suficiente para gerar probabilidades para as palavras e consequentemente erros no reconhecimento. Todos os modelos de linguagem gerados nos experimentos foram avaliados com textos de videoaulas pois era este o contexto de aplicação de reconhecimento de fala que está sendo avaliado. A Tabela 4.2 apresenta os resultados dos experimentos com modelo de linguagem.

Tabela 4.2: Resultados para modelo de linguagem

	Perplexidade	
Versão	**3-gram**	**4-gram**
LM1	642.156	642.156
LM2	471.256	471.256
LM3	382.473	382.473
LM4	368.504	341.637
LM5	364.193	**335.880**

A principal diferença nos valores da perplexidade está relacionado a inserção de texto ao *corpus*. Isto se deve, principalmente, porque há um aumento da abrangência do vocabulário e com isso diminui a chance de não existir uma palavra dentro do vocabulário.

A normalização também influenciou positivamente no resultado. Isto se deve pois alguns lixos presentes no *corpus* (ex: pontuações, caracteres especiais etc) foram excluídos e a expansão dos padrões tornou a base mais uniforme com relação a esses termos (ex: dois e 2).

Os ganhos obtidos nos outros experimentos foram menos significativos porém, diferentes métodos de descontos fizeram diferenças interessantes nos modelos criados. Isto não quer dizer que o método de desconto utilizado no LM5 é o melhor método da literatura, e sim que é o melhor para a base de avaliação utilizada e consequentemente para o contexto em que o modelo de linguagem será utilizado.

Apesar de ganhos expressivos com os experimentos, os valores de perplexidade do melhor modelo de linguagem para a base de avaliação ainda é alto. Isto é decorrente principalmente da característica da base de avaliação. O texto de videoaulas, geralmente, é muito informal e sem uma ordem de escrita e de assunto. Outro fator importante para este valor de perplexidade alto é o contexto de construção do *corpus* de texto, pois a grande maioria das frases é oriunda de textos do Wikipédia e da base do CETEN que são textos formais e escritos em norma culta da língua portuguesa. Contudo, para a aplicação proposta, seria o modelo ideal.

4.3 Resultados do reconhecimento de fala

O modelo acústico será avaliado em um sistema de reconhecimento automático de fala. Este sistema recebe o modelo acústico gerado e decodifica o áudio em texto a partir deste modelo. Algumas versões de modelos de linguagem, gerados nos experimentos, serão utilizados em conjunto com o modelo acústico a fim de criar uma estrutura de sistema de reconhecimento de fala.

Todas as possíveis combinações de modelos acústicos e modelos de linguagem não foram geradas, pois o tempo de compilação de ambos os modelos em um sistema de reconhecimento de fala e o tempo gasto no processamento da base de avaliação é elevado. Contudo, a combinação de modelos avaliados já representa uma tendência de melhora dos resultados do reconhecimento de fala. Foi avaliado, principalmente, o efeito do processo de normalização no texto do treinamento do modelo de linguagem. As combinações de modelos avaliam as opções de modelos acústicos gerados e a relevância do processo de normalização do modelo de linguagem no reconhecimento de fala. A Tabela 4.3 apresenta o resultados da taxa de WER (*Word Error Rate*) para a combinação de modelos avaliados. Quanto menor a taxa de WER, mais correto é o reconhecimento.

Tabela 4.3: Resultados do modelo acústico e modelo de linguagem

id	**Experimento**	**WER (%)**
1	AM1 + LM2	94,60
2	AM2 + LM2	92,00
3	AM3 + LM5	48,10
4	AM4 + LM2	54,30
5	AM4 + LM5	**45,50**

A diferença entre os experimentos 1 e 2 representa as melhorias encontradas no uso da técnica de DNN em comparativo com o modelo acústico GMM. O alto valor de WER está relacionado principalmente a ausência do método de alinhamento proposto que, consequentemente, influenciou no treinamento do modelo acústico com o *corpus* de voz mal alinhado.

O principal motivo da queda na taxa de WER do experimento 2 para o experi-

mento 3, 4 e 5 é com relação ao método de alinhamento que foi proposto. Esta diferença não quer dizer que a legenda das videoaulas estavam incorretas com relação ao tempo do video, e sim que para uma aplicação de treinamento de modelos acústicos é necessário que este alinhamento esteja o mais correto possível. Quando uma pessoa está interagindo com uma videoaula através da legenda, este erro de alinhamento passa despercebido e não causa nenhum incômodo. O principal problema de um alinhamento errado é que o algoritmo que gera os modelos atribui a um vetor de características acústicas uma palavra ou uma frase que não reflete o que está sendo dito naquele áudio, gerando assim probabilidades erradas.

A queda da taxa de WER do experimento 4 para o 5 se deve exclusivamente pelo modelo de linguagem normalizado que foi utilizado no experimento 5. Esta melhora na WER é importante para um sistema de reconhecimento de fala e representa que este sistema está mais robusto e correto em relação a fala humana e a língua portuguesa. O modelo de linguagem, neste contexto, auxilia o sistema para que sejam retornadas frases corretas gramaticalmente e não apenas palavras soltas.

Por fim, a diferença entre os experimentos 3 e 5 é referente as configurações de áudio, que foram alteradas nestas duas versões de modelo. Este experimento se deve, principalmente, por uma tendência encontrada nas videoaulas que foram usadas para avaliar os modelos. É importante que o modelo gerado tivesse características acústicas bem parecidas com as características dos áudios da base de avaliação, pois assim a chance de encontrar alguma equivalência na modelagem é maior. O ganho nesta etapa foi menor, porém para estes sistemas de reconhecimento, qualquer ganho de WER é significativo, principalmente quando este sistema já é mais próximo do sistema ideal (WER = 0).

4.4 Comparativo com sistemas comerciais

A fim de ter algumas métricas de referência para os experimentos propostos, mediu-se a taxa de erros de palavras de 3 sistemas comerciais (Google[1], IBM[2] e Microsoft[3]) sobre as

[1] `https://cloud.google.com/speech/`

[2] `https://www.ibm.com/watson/developercloud/speech-to-text.html`

[3] `https://docs.microsoft.com/pt-br/azure/cognitive-services/Speech/API-Reference-REST/BingVoiceRecognition`

bases de teste em comparação com o melhor resultado de experimento de reconhecimento de fala. A mesma base de avaliação foi submetida a esses sistemas e o resultado está representado na Tabela 4.4.

Tabela 4.4: WER de Sistemas Comerciais

	WER (%)
AM4 + LM5	45,50
Google	35,90
IBM	73,70
Microsoft	44,70

O melhor resultado dos experimentos deste trabalho tem um desempenho melhor que o sistema da IBM, fica bem próximo do sistema da Microsoft e perde para o Google. Os modelos gerados pelo experimento tem a vantagem de terem sido treinados com áudios mais próximos da aplicação alvo (o que tende a ser uma vantagem em relação aos sistemas comerciais). Por outro lado, a quantidade de material usado nos treinamentos deste trabalho é muito menor que aquela usada pelos sistemas comerciais. Enquanto Google, Microsoft e IBM devem juntar algumas milhares de horas, o treinamento feito no experimento AM4 + LM5 conta com apenas algumas dezenas de horas.

Vale ressaltar que os números levantados para os sistemas comerciais foram medidos durante o primeiro trimestre de 2017 e usaram as chamadas padrão das APIs disponibilizadas pelos fabricantes. Obviamente, os números poderão ser diferentes se o teste for repetido com parâmetros diferentes nas chamadas das APIs ou depois que algum fabricante implementar melhorias em seu sistema.

4.5 Análise de Resultados

A principal medida utilizada para avaliar os modelos de reconhecimento de fala é a taxa de erro de palavras (WER). Como o objetivo de investigar as tarefas com melhor custo benefício podemos concluir que o método de alinhamento da base proposto, apresentou o melhor ganho com relação a WER e tornou os modelos competitivos com relação a sistemas comerciais. O gráfico representado pela Figura 4.1 apresenta um relação entre o

tempo gasto para gerar cada experimento (desde a etapa de criação da base de treinamento até a avaliação dos modelos gerados) e a WER com relação a base de avaliação.

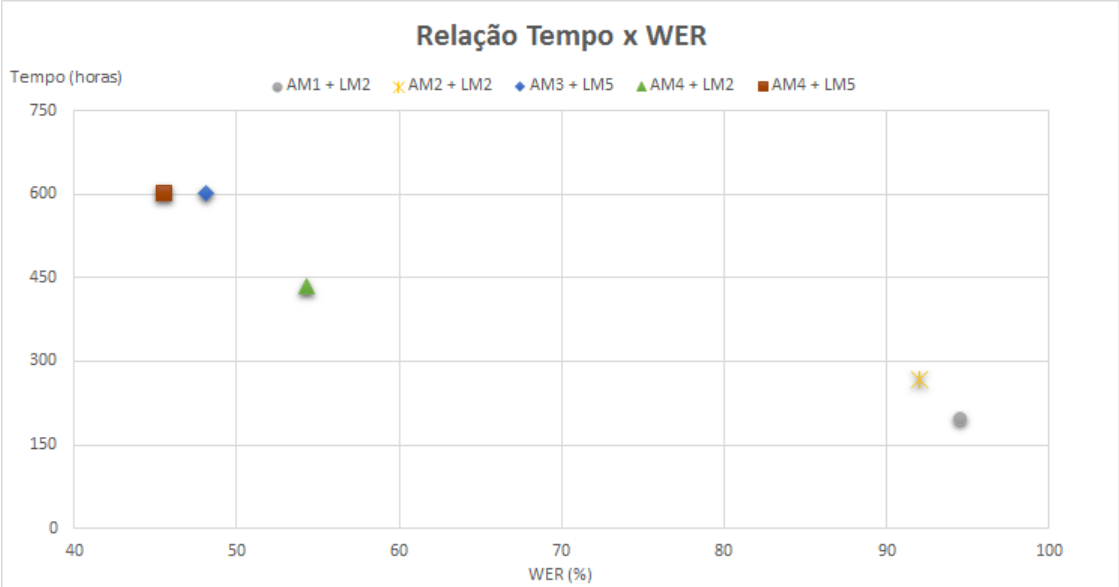

Figura 4.1: Gráfico da Relação Tempo x WER

O eixo X neste gráfico representa a taxa de WER. Como nosso objetivo é diminuir esta taxa, os modelos que se aproximem de 0 serão os melhores. O eixo Y representa o tempo gasto neste experimento. Este tempo é a soma do tempo de todas as etapas gastas na construção de um modelo acústico e de um modelo de linguagem.

Pode-se observar que o experimento AM4 + LM5 e o experimento AM3 + LM5 possuem o mesmo tempo para criação e a WER tem uma pequena melhora. Esta melhora está relacionada as alterações efetuadas nas configurações de áudio do *corpus* de voz. Assim, esta etapa possui um bom custo beneficio e, por mais que o ganho tenha sido pequeno, quando se trata de WER, toda melhora nesta taxa é significativo no resultado final.

A melhora na WER do experimento AM4 + LM5 para o experimento AM3 +LM5 é decorrente do processo de normalização do *corpus* de texto. O tempo decorrente desta melhora é alto, porém é importante ressaltar que a análise de tempo não quer dizer que este recurso é escasso e deve ser minimizado ao máximo, mas sim que por mais que o tempo gasto nesta tarefa seja alto o ganho também é. Este trabalho tem como objetivo orientar o gasto de tempo em tarefas que valem a pena e não diminuir o tempo do processo de treinamento de sistemas de reconhecimento de fala. Isto quer dizer que vale a pena o

empenho nesta tarefa pois o resultado é satisfatório.

A melhora mais significativa com relação ao tempo e a WER está relacionado aos experimentos AM2 + LM2 e AM4 + LM2. Esta melhora está relacionada ao alinhamento do *corpus* de voz. Está tarefa, é a melhor com relação ao custo beneficio pois foi a que obteve o melhor desempenho nos experimentos.

Por fim, pode-se concluir que o experimento AM4 + LM5, que obteve a menor WER, foi o melhor em comparação com os outros e que, por mais que o tempo gasto nas tarefas deste experimento seja alto, vale a pena este gasto de tempo pois os ganhos obtidos foram altos e permitiram o uso desses modelos de forma robusta em um sistema de reconhecimento de fala.

5 Conclusões

A partir dos resultados apresentados podemos concluir, das etapas investigadas do processo de treinamento de modelos acústicos e de linguagem, quais são críticas e mais influenciam na acurácia do reconhecimento de fala.

As melhorias propostas no *corpus* de voz e texto foram os principais elementos que influenciaram nos ganhos obtidos no reconhecimento automático de fala. As etapas de alinhamento do áudio e normalização do texto trouxeram ganhos expressivos na acurácia dos sistemas de reconhecimento de fala avaliados. O tempo de criação dos algoritmos para alinhamento e normalização foi de 7 dias cada, porém, a execução destes algoritmos não possui tempo elevado e o ganho nestas tarefas foi o mais expressivo. Assim, pode-se concluir que a base de dados utilizada no treinamento é o principal elemento de influência nos resultados de sistemas de reconhecimento de fala e possui o melhor custo benefício das tarefas avaliadas. Isto se deve pois as técnicas de treinamento de modelos acústicos e de linguagem já são métodos eficientes que provém de muito esforço da literatura na pesquisa por reconhecimento de fala. Com isso, o principal efeito na acurácia destes sistemas está relacionado a base de treinamento, sua abrangência e o contexto de aplicação.

Este trabalho trás como contribuição a criação de um *corpus* de voz[1] gratuito para o português e instruções para processamento e melhor aproveitamento deste *corpus* num contexto de reconhecimento de fala. Na literatura não foram encontradas bases de dados gratuitas e com dimensões razoáveis como é proposto neste trabalho.

5.1 Limitações

Os experimentos desenvolvidos pelo trabalho tornaram possível o entendimento das etapas criticas no processo de treinamento de modelos para sistemas de reconhecimento de fala, porém estas conclusões não podem ser expandidas para todos os contextos de aplicações e de criação de sistemas de reconhecimento de fala. Em contextos extremamente diferentes,

[1] `https://goo.gl/U3eTLk`

são necessários novos experimentos e gasto de tempo em novas tarefas, que não foram abordadas neste trabalho.

Outra limitação é com relação a quantidade de combinações de modelos acústicos e de linguagem que foram avaliados. Por mais que os resultados pareçam ser conclusivos a partir dos experimentos, a geração de outras combinações podem apresentar melhor o resultado de melhoria das diversas tarefas.

Por fim, não foi avaliado como o aumento do *corpus* de voz influencia na acurácia do reconhecimento de fala.

5.2 Trabalho Futuros

Como trabalho futuro, pretende-se principalmente aumentar a *corpus* de voz através de *audiobooks*(Panayotov et al., 2015) e criar um método para alinhamento forçado do texto desses livros ao áudio disponibilizado. Ainda, pretende-se avaliar com isso a interferência do aumento do *corpus* de voz na acurácia do sistema de reconhecimento de fala.

Com relação ao *corpus* de texto, pretende-se também aumentar o número de sentenças através de *datasets* de texto disponibilizados na WEB, criar um algoritmo para correção ortográfica de palavra e observar novos padrões de texto para serem expandidos em sequências textuais.

Por fim, com relação ao métodos de treinamento de modelos, pretende-se continuar avaliando as diversas formas de treinamento de modelos acústicos que são disponibilizados na literatura e quais são os novos métodos que apresentam melhores resultados.

Bibliografia

Campos, V. P.; Araujo, T. ; Souza Filho, G. Cinead: Um sistema de geração automática de roteiros de audiodescrição. **Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)**, 2014.

Chan, H. Y.; Woodland, P. C. **Improving broadcast news transcription by lightly supervised discriminative training**. In: In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p. 737–740, 2004.

Chen, S. F.; Goodman, J. **An empirical study of smoothing techniques for language modeling**. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics, p. 310–318. Association for Computational Linguistics, 1996.

Carvalho, B. B.; Souza, J. F. d. Anotação semântica de transcritos para indexação e busca de vídeos. **Relatórios Técnicos do DCC/UFJF**, 2017.

Cuadros, C. D. R.; FERREIRA, E. L. C. O. Reconhecimento de voz e de locutor em ambientes ruidosos: comparação das técnicas mfcc e zcpa. **Programa de Pós-graduação em Engenharia de Telecomunicações**, 2007.

Evermann, G.; Chan, H. Y.; Gales, M. J.; Jia, B.; Mrva, D.; Woodland, P. C. ; Yu, K. **Training lvcsr systems on thousands of hours of data.** In: ICASSP (1), p. 209–212, 2005.

Developers, F. **ffmpeg tool (version be1d324)** [**software**]. url http://ffmpeg.org/, 2017.

Ferreira, M. V. G.; de Souza, J. F. **Use of automatic speech recognition systems for multimedia applications**. In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web, p. 33–36. ACM, 2017.

Ferreira, M. V. G.; Souza, J. F. d. **Utilização de sistemas para reconhecimento automático de fala para aplicações multimídia**. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Minicursos, 2017, 139-178p.

Gonzalez, M.; Lima, V. L. S. **Recuperação de informação e processamento da linguagem natural**. In: XXIII Congresso da Sociedade Brasileira de Computaçao, volume 3, p. 347–395, 2003.

Higgins, E. L.; Raskind, M. H. Speaking to read: The effects of continuous vs. discrete speech recognition systems on the reading and spelling of children with learning disabilities. **Journal of Special Education Technology**, v.15, n.1, p. 19–30, 1999.

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

Hämäläinen, A.; Pinto, F. M.; Rodrigues, S.; Júdice, A.; Silva, S. M.; Calado, A. ; Dias, M. S. **A multimodal educational game for 3-10-year-old children: Collecting and automatically recognising european portuguese children's speech**. In: Speech and Language Technology in Education, 2013.

Katz, S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. **IEEE transactions on acoustics, speech, and signal processing**, v.35, n.3, p. 400–401, 1987.

Ma, J.; Matsoukas, S.; Kimball, O. ; Schwartz, R. **Unsupervised training on large amounts of broadcast news data**. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, volume 3, p. III–III. IEEE, 2006.

Meinedo, H.; Caseiro, D.; Neto, J. ; Trancoso, I. Audimus. media: a broadcast news speech recognition system for the european portuguese language. **Computational Processing of the Portuguese Language**, p. 196–196, 2003.

Neto, N.; Silva, Ê. ; Sousa, E. **Software usando reconhecimento e síntese de voz: o estado da arte para o português brasileiro**. In: Proceedings of the 2005 Latin American conference on Human-computer interaction, p. 326–331. ACM, 2005.

Ney, H.; Essen, U. **On smoothing techniques for bigram-based natural language modelling**. In: Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on, p. 825–828. IEEE, 1991.

Oliveira, R.; Batista, P.; Neto, N. ; Klautau, A. Recursos para desenvolvimento de aplicativos com suporte a reconhecimento de voz para desktop e sistemas embarcados. **12o Fórum Internacional de Software Livre**, 2011.

Panayotov, V.; Chen, G.; Povey, D. ; Khudanpur, S. **Librispeech: an asr corpus based on public domain audio books**. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 5206–5210. IEEE, 2015.

Pessoa, L. S.; Violaro, F. ; Barbosa, P. A. Modelos da lingua baseados em classes de palavras para sistema de reconhecimento de fala continua. **Revista da Sociedade Brasileira de Telecomunicações**, v.14, n.2, p. 75–84, 1999.

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P. ; others. **The kaldi speech recognition toolkit**. In: IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011.

Rabiner, L.; Juang, B. An introduction to hidden markov models. **IEEE assp magazine**, v.3, n.1, p. 4–16, 1986.

Raimond, Y.; Lowis, C. Automated interlinking of speech radio archives. **LDOW**, v.937, 2012.

Riccardi, G.; Hakkani-Tür, D. Z. **Active and unsupervised learning for automatic speech recognition.** In: INTERSPEECH, 2003.

Silva, E.; Pantoja, M.; Celidônio, J. ; Klautau, A. **Modelos de linguagem n-grama para reconhecimento de voz com grande vocabulário**. In: III workshop em tecnologia da informação e da linguagem humana, 2004.

Ênio Silva; Baptista, L.; Fernandes, H. ; Klautau, A. **Desenvolvimento de um sistema de reconhecimento automático de voz contınua com grande vocabulário para o português brasileiro**. In: XXV congresso da sociedade Brasileira de computação, 2005.

Silva, C. P. A. d. **Um software de reconhecimento de voz para português brasileiro**. Dissertação de Mestrado - .

Siravenha, A.; Neto, N.; Macedo, V. ; Klautau, A. **Uso de regras fonológicas com determinaçao de vogal tônica para conversao grafema-fone em português brasileiro**. In: 7th International Information and Telecommunication Technologies Symposium, p. 1–6, 2008.

Stolcke, A.; others. **Srilm-an extensible language modeling toolkit.** In: Interspeech, volume 2002, p. 2002, 2002.

Tevah, R. T. **Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro**. 2006. Tese de Doutorado - UNIVERSIDADE FEDERAL DO RIO DE JANEIRO.

Thomas, S.; Seltzer, M. L.; Church, K. ; Hermansky, H. **Deep neural network features and semi-supervised training for low resource speech recognition**. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, p. 6704–6708. IEEE, 2013.

Tur, G.; Stolcke, A.; Voss, L.; Dowding, J.; Favre, B.; Fernández, R.; Frampton, M.; Frandsen, M.; Frederickson, C.; Graciarena, M. ; others. **The calo meeting speech recognition and understanding system**. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE, p. 69–72. IEEE, 2008.

da Veiga, A. O. **Treino não supervisionado de modelos acústicos para reconhecimento de fala**. 2013. Tese de Doutorado - Universidade de Coimbra.

Wessel, F.; Ney, H. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. **IEEE Transactions on Speech and Audio Processing**, v.13, n.1, p. 23–31, 2005.

Yang, H.; Meinel, C. Content based lecture video retrieval using speech and video text information. **IEEE Transactions on Learning Technologies**, v.7, n.2, p. 142–154, 2014.

Ynoguti, C. A. **Reconhecimento de fala contínua usando modelos ocultos de Markov**. 1999. Tese de Doutorado - Universidade Estadual de Campinas.

Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D. ; others. The htk book (v3. 4). **Cambridge University**, 2006.