Síntese de Fala com Redes Neurais

Marcos Pedro Ferreira Leal Silva

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Instituto de Computação
Orientador: Dr. Marcelo Queiroz

São Paulo, fevereiro de 2019

Síntese de Fala com Redes Neurais

Esta é a versão original da dissertação elaborada pelo candidato (Marcos Pedro Ferreira Leal Silva), tal como submetida à Comissão Julgadora.

Resumo

Leal, M., **Síntese de Fala com Redes Neurais**. 2018. 28 f. Tese (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

O presente trabalho apresenta a aplicação de redes neurais artificiais profundas interconectadas de modo a sermos capaz de modelar fala e reproduzi-la com trechos que nunca foram falados pelo falantes original. Para esse fim exploramos na revisão bibliográfica os algoritmos que atingiram o estado da arte da síntese de fala e oferecemos uma análise comparativa entre as estratégias adotadas. Criticamos a majoritariedade de modelos desenvolvidos apenas com a linguagem inglesa e apontamos algumas diferenças linguísticas entre o português e o inglês. Apresentamos então a implementação de um desses modelos para validar sua performance na língua portuguesa. Comparamos os resultados obtidos com os já publicados para levantar um comparativo quanto à performance do modelo em tempo de execução/treino e também quanto à percepção humana sobre a voz sintetizada através do *MOS (Mean Opinion Score)*.

Palavras-chave: redes neurais; síntese; fala; síntese de fala neural;

Abstract

LEAL, M. **Voice Synthesis with Deep Neural Networks**. 2019. 28 f. Tese (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

The paper presents the application of deep neural networks in order to model speech and reproduce it with excerpts that were never spoken by the original speakers. For this we explored in the bibliographic review the algorithms that reached the state of the art of speech synthesis and offered a comparative analysis between the adopted strategies. We criticize the majority of models developed only with the English language and point out some linguistic differences between Portuguese and English. We then present the implementation of one of these models to validate their performance in the Portuguese language. We compared the results obtained with those already published in order to compare the performance of the model at runtime / training and also the human perception about the voice synthesized through MOS (Mean Opinion Score).

Keywords: neural networks; synthesis; speech; neural speech synthesis;

Contents

Lista de Abreviaturas vii

List of Figures ix

List of Tables xi

1 Introdução 1
- 1.1 Organização do Trabalho . . . 1
- 1.2 Contribuições . . . 2

2 Conceitos 5
- 2.1 Motivação . . . 5
- 2.2 Histórico . . . 5
  - 2.2.1 1790 - von Kempelen e Wheatstone . . . 6
  - 2.2.2 1939 - Homer Dudley e o Voder . . . 6
  - 2.2.3 1970 - Gunnar Fant e a Síntese Articulatória . . . 7
  - 2.2.4 Síntese Concatenativa . . . 8
  - 2.2.5 Síntese de Formantes . . . 10
  - 2.2.6 Aplicações e Interesse Comercial . . . 10
- 2.3 Alfabetos Fonéticos . . . 12
- 2.4 Redes Neurais e Aprendizado de Representações . . . 15
  - 2.4.1 Representações Densas . . . 15
  - 2.4.2 RNN - Redes Neurais Recorrentes . . . 16
  - 2.4.3 Conectist Temporal Classification (CTC) e Beam Search . . . 17
  - 2.4.4 Convoluções Autoregressivas Dilatadas . . . 18

3 Metodologia 19
- 3.1 Construção do Dataset . . . 19
  - 3.1.1 Álinhamento de Legendas do YouTube . . . 20

4 Cronograma 21
- 4.1 Cronograma Previsto . . . 21

A 23
- A.1 Tabela IPA . . . 24

Bibliography 25

Index **28**

Lista de Abreviaturas

CNN	Redes Neurais Convolucionais (*Convolutional Neural Netowkrs*)
RNN	Redes Neurais Recorrentes (*Recurrent Neural Networks*)
TTS	Texto para Fala (*Text to Speech*)
LSTM	Memória de Longo-Curto Prazo (*Long Short Term Memory*)
GRU	Unidade Recorrente Bloqueada (*Gated Recurrent Units*)
STT	Fala para Texto (*Speech to Text*)

List of Figures

1.1 Diagrama do Processo de Produção/Percepção da Fala 2

2.1 Máquina de von Kempelen-Wheatstone 6
2.2 Voder apresentador em 1939 7
2.3 Espectrograma da frase “greetings everybody” 8
2.4 Esquema de Filtragem de um *Voder* 9
2.5 Alfabeto de Hiragana com Fonético Ocidental Respectivo 10
2.6 Blocos de Síntese do Deep Voice 12
2.7 Correlação entre Datasets 13
2.8 Distribuição de Fonemas por conjunto de dados 14
2.9 Célula Recorrente 16
2.10 Células Recorrentes 17
2.11 Convolução Dilatada 18

List of Tables

Chapter 1

Introdução

A fala é a forma mais natural de comunicação que uma pessoa tem no cotidiano. Bem recentemente nossos computadores têm se tornado cada vez melhores na complexa tarefa que é interagir conosco usando a fala. Não é claro quando os humanos começaram a falar propriamente uns com os outros mas através do nosso longo período de evolução fomos adquirindo maneiras mais precisas, específicas e claras de expressar e comunicar ideias e transmitir informações através de som com o uso de uma linguagem.

O processo todo de interação engloba uma sequência de atividades (Fig. 1.1): a organização de uma ideia em palavras, a expressão dessas palavras por um meio vocal, a transmissão desse meio até o aparelho auditivo do par que se interage, a conversão das onda sonora pelo aparelho auditivo do ouvinte para sinais elétricos transmitidos e processados pelo cérebro, o processamento desses sinais em uma linguagem e finalmente a interpretação desses sinais em um significado completo.

Sistemas de STT (*Speech to Text*) foca no processo D representado no diagrama, onde o interesse é apenas a compreensão da mensagem transmitida. O processo C de transmissão no meio é interpretado através de um processo de quantização e amostragem do contínuo de modo a ser possível seu processamento por computadores. O papel de formulação e compreensão da mensagem transmitida representado no diagrama pelas letras A e E, respectivamente, já teve vários candidatos ao longo dos anos, desde sistemas especialistas, diálogos com chatbots até complexas redes neurais para respostas de questões genéricas (**? ? ? ? ? ?**). O intuito desse trabalho é o estudo da replicação do modelo de fala, representado pela letra B, e pela síntese e transmissão do mesmo por um meio, a letra C. Não é intuito desse trabalho dar ênfase a nenhum dos outros momentos de produção, percepção e interação através da fala mas dada uma intercessão natural existente entre as atividades de comunicação alguns cuidados e algumas técnicas são compartilhadas entre processos.

As primeiras tentativas documentadas de estratégias para síntese de fala datam do século XVIII (VK91). As mais diversas técnicas foram tomando forma e evoluindo com o tempo de modo que, comparativamente, algumas são mais simples, outras mais portáteis, outras mais genéricas, todas visando algum aspecto distinto cujo algum outro modelo era ineficiente para a aplicação desejada. Abordo com ênfase nesse trabalho a síntese de fala neural que é uma das estratégias mais genéricas dentre as disponíveis atualmente pois permite capturar as características necessárias para a síntese nas mais diversas fontes e dos mais diversos falantes.

1.1 Organização do Trabalho

Começamos esse trabalhos contextualizando técnicas implementadas anteriormente e técnicas usadas amplamente ainda hoje, caracterizando assim um breve revisão histórica das abordagens para síntese de fala. Após contextualizar os modelos legados iremos apontar os atuais estados da arte para síntese de fala neural focando na sua construção modular e discutir seus aspectos positivos e negativos comparativamente. Após introduzir os modelos neurais estudados desejamos dar ao leitor uma introdução dos conceitos utilizados em cada modelo.

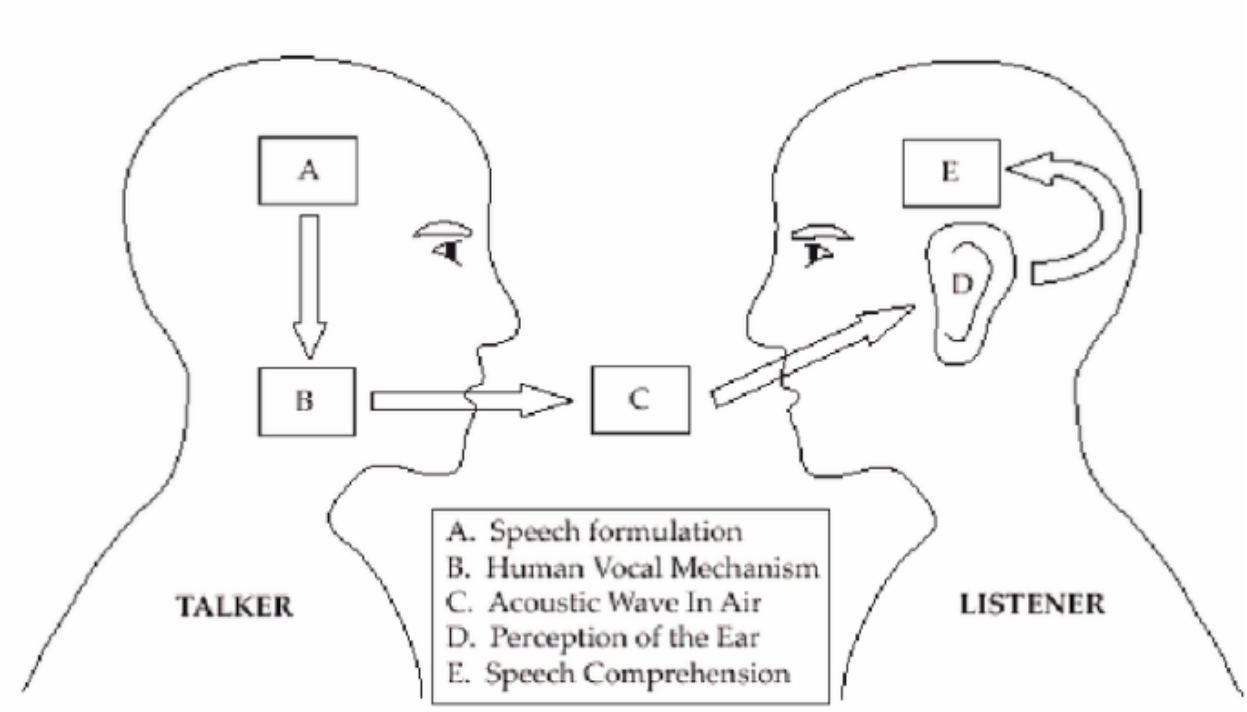


Figure 1.1: *Diagrama do Processo de Produção/Percepção da Fala (Tun05)*

Selecionamos então um modelo neural para discorrer com maiores detalhes e buscamos estudar seu comportamento com a língua portuguesa. Iniciamos a discussão através da análise dos dados e de um possível enviesamento de performance do modelos tendo em vista a majoritariedade das análise focando apenas a síntese no inglês. Levantamos um estudo observando diferenças entre português e inglês como diversidade e frequência de fonemas, sotaques, dimensão de palavras existentes, dimensão de palavras usadas cotidianamente. Tendo essa análise comparativa entre os dados de treinamento esperamos ter mais facilidade em criticar o comportamento do modelo tanto no inglês quanto no português.

Apontamos uma estratégia genérica para construção de um conjunto de dados com frases/-palavras curtas em português através da extração de áudios de vídeos legendados. Usamos nesse trabalho áudios de vídeos do YouTube baseado nas segmentações automáticas das legendas. Apontamos as vantagens e desvantagens desse método e como parâmetro comparativo elegemos uma situação de ilusão auditiva (Mas07) para observar se o modelo é capaz de produzir alguma síntese similar.

Discutimos então os detalhes técnicos de implementação, disponibilizamos os parâmetros utilizados no treinamento, discutimos as métricas usadas para avaliação da performance do modelo durante o treino e então sintetizamos algumas frases referenciais para poder avaliar a performance do sistema na avaliação humana. Essas frases selecionadas sintetizadas são avaliadas por um grupo de ouvintes que dão notas cujos valores mais altos correspondem aos trechos mais realistas.

1.2 Contribuições

Buscamos nesse trabalho:

1. Propor uma estratégia de construção escalável e automática de uma base de dados não-estruturados disponíveis em alta escala para qualquer língua

2. Avaliar a performance do estado da arte de síntese de fala com estratégias usando redes neurais na língua portuguesa comparativamente ao inglês

3. Levantar aspectos interpretáveis do aprendizado do modelo neural de modo a facilitar possíveis otimizações futuras

4. Propor um modelo de *embeddings* que faça jus à similaridade fonética das palavras de modo que na conversão grafema → fonema possamos ter uma contextualização mais próxima à falada possível

Chapter 2

Conceitos

Abordamos nesse capítulo as técnicas de síntese desenvolvidas historicamente até o alcance da síntese neural, foco desse trabalho. O intuito de discutir a evolução de técnicas históricas é observar o desenvolvimento de possíveis estratégias que possam ser permutáveis entre técnicas e quais preocupações buscavam resolver. Como comentado anteriormente, na discussão de outras técnicas acabamos permeando outras áreas de pesquisa com intercessões nos assuntos que nos interessam mas que permeiam outros objetivos em suas respectivas áreas. Esses tópicos, em particular, são abordados de maneira superficial já que existem motivos próprios para decisões dentro de cada atividade e, ciente dessas possibilidades, podemos alinhar respectivamente os modelos para interação ótima. Trazemos os modelos mais recentes de síntese de fala neural e discutimos as diferenças entre a **adaptação de falante** e o **aprendizado de parâmetros para cada falante**, que são as estratégias mais populares. Trago outras estratégias possíveis dentro do universo de síntese neural apresentadas em paralelo à solução implementada neste trabalho.

2.1 Motivação

A síntese de fala a partir do texto (TTS) pode ser utilizada em diversas aplicações como dispositivos como dispositivos com fala embutida, sistemas de navegação e auxílio para as pessoas com problemas de visão. Na sua natureza mais simples a interação com voz é uma das primeiras formas de comunicação complexa que aprendemos e somos capazes de expressar interações avançadas quando comparada a formas mais primitivas da infância. A fala permite, essencialmente, a interação sem a necessidade de interação visual.

Sistemas de TTS atualmente são modelos complexos baseados em uma cadeia de processos com vários estágios onde existem vários aspectos e características controlados manualmente e através de heurísticas. Fruto dessa cadeia complexa temos sistemas cuja natureza intrínseca é trabalhosa e de difícil compreensão. Os modelos neurais também seguem uma cadeia similar aos modelos clássicos mas ao invés de termos heurísticas e parâmetros manualmente alocados buscamos através das redes os parâmetros que otimizam o dado problema.

2.2 Histórico

O ser humano desde o início da sua evolução adotou diferentes maneiras para transmitir informações entre os mesmos de sua espécie. Um dos métodos mais primordiais que é compartilhado por diversas outras espécies é a capacidade de produzir sons e atribuir significados a eles. Diversas espécies possuem sons específicos para alertas sobre um possível predador ou para acasalar e com os seres humanos não é diferente.

Uma das fontes de produção de sons na nossa espécie é a vibração das pregas vocais localizadas na laringe. A contração ou relaxamento das pregas é responsável por gerar diferentes sons quando cortadas por um fluxo de ar. Outro fator importante para a produção de som é a ressonância gerada nas cavidades do corpo, especialmente a bucal e nasal. Desse modo podemos produzir

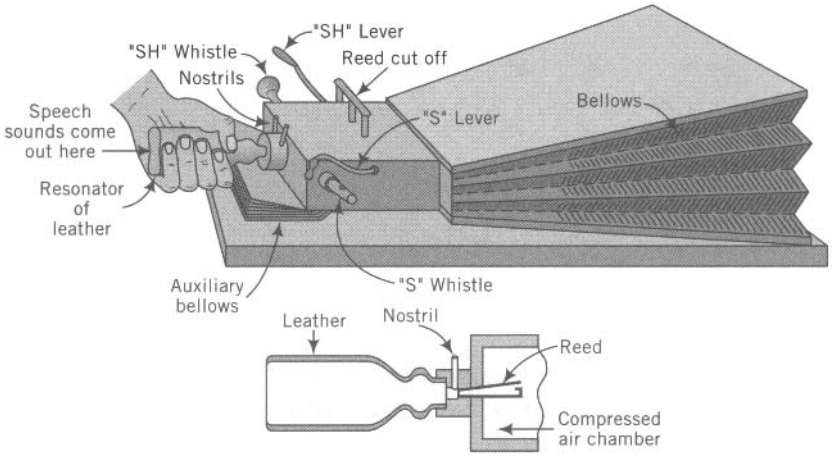


Figure 2.1: *Máquina de von Kempelen-Wheatstone*

diferentes sons com a boca mais ou menos aberta, sons nasalados ou não. Por fim temos nossa língua e lábios que permitem produzir uma miríade de sons através dos mais diversos movimentos e estalos, por exemplo. Todo esse aparato permite que a voz humana seja um complexo mecanismo de amplificação e filtragem nos possibilitando a fala, o riso, o grito e a produção de outros sons diversos.

A capacidade de replicar sons é observada em várias situações na natureza. Podemos apontar espécies como o pássaro de Lira ou o papagaio que são capazes de repetir algum som ouvido ou mesmo na nossa espécie, onde algumas pessoas com excelente controle do aparelho vocal são capazes de adaptá-lo e simular a voz de outras pessoas. Nossa capacidade de armazenar, reproduzir e sintetizar sons começou a ser amplamente possível apenas no final do século XIX. Até esse trecho da história muito havia sido aprendido sobre sons, diversos instrumentos musicais haviam sido criados, muito da questão física do som como onda fora estudado e com todo esse conhecimento que fomos capazes de rapidamente desenvolver diversas tecnologias para trabalhar com som.

2.2.1 1790 - von Kempelen e Wheatstone

O interesse na síntese de voz teve seus primeiros resultados com von Kempelen em 1790. Em seu livro (VK91) ele retrata os aspectos da origem da fala, do sistema de produção de fala nos seres humanos e sobre sua máquina de fala. Uma possível interpretação do esquema sugerido pode ser visto na figura 2.1 como foi construído por Wheatstone anos depois. Podemos perceber a presença de diversas chaves e alavancas para simular diversos fonemas, a presença de um pulmão artificial pelo fole e uma câmara de couro funcionando como um simples ressonador. Esse modelo é capaz de reproduzir apenas alguns fonemas e sua complexa operação nunca permitiu a síntese de sentenças complexas.

2.2.2 1939 - Homer Dudley e o Voder

Anos mais tarde um pesquisador da *Bell Telephone Laboratory*, Homer Dudley, buscou continuar a análise de von Kempelen em seus trabalhos (DT50) (GME11) agora com os aparatos eletrônicos disponíveis. Em 1939 ele chamou a atenção do mundo ao apresentar na Grande Feira Mundial de 1939 de São Francisco e Nova Iorque o dispositivo chamado de *Voder* (***Vo****ice-operated* ***De****monstrator*). Esse dispositivo que pode ser visto na figura 2.2 produzia sons a partir dos movimentos do usuário sobre as pedaleiras, manivelas e ajustes que filtravam uma fonte de ruído, apresentando resultados como na síntese subtrativa. O *Voder* não era capaz de produzir sons sem um habilidoso operador dada sua grande quantidade de parâmetros a serem controlados simultaneamente. Muitos dos operadores iniciados no treinamento não foram capazes de operar ou demoraram até um ano para conseguir manipular com destreza o aparelho.

Comparando o espectrograma de um falante humano e de de um áudio gerado por um *Voder* (Fig. 2.3) percebemos que a síntese ainda não era capaz de capturar toda a riqueza do espectro humano especialmente nas harmônicas e frequências mais altas. O espectrograma é uma das ferramentas mais poderosas para observar o comportamento dos sintetizadores e comparar com a voz humana. Percebemos também que como o *Voder* por ser operado por controles manuais apresentava fonemas naturalmente mais longos de modo que o operador fosse capaz de executá-los, outro fator para o forte sentimento sintético desse som.

No início do século XX temos uma forte influência de Alan Turing para o interesse pela síntese de voz devido ao seu famoso Teste de Turing. A fala é um fator importante no teste de Turing completo onde haveria uma interação total entre a máquina e o usuário através de um robô. Nesse aspecto a fala deveria ser indistinguível de um falante humano para nos confundir perfeitamente. Turing apresentou sua tese em 1950 e alguns anos depois já existiam alguns projetos para síntese existentes. A ideia geral de diversos projetos pode ser observada na figura 2.4 (inclusive do próprio *Voder* de Dudley) onde uma fonte de ruído é filtrada e ressoada de acordo com o interesse do usuário em produzir diferentes fonemas.

Em 1961 ainda na *Bell Labs Technology* a pesquisa na síntese de fala iniciada por Dudley continuava e John Larry Kelly Jr. juntamente com Louis Gerstman usaram um IBM 704 para síntese. Um episódio popularmente conhecido foi a síntese de "Daisy Bell" gravada por eles ter sido ouvida por Arthur C. Clarke que visitava um colega. Arthur ficou tão impressionado com a tecnologia de fala que acabou tornando-se parte de sua obra *2001: Uma Odisseia no Espaço* através do robô HAL 9000 [1].

2.2.3 1970 - Gunnar Fant e a Síntese Articulatória

Em 1970, Gunnar Fant publicou seu livro *Teoria Acústica da Produção da Fala* (Fan60) e nele defendeu um modelo de tubos para modelar a voz humana. Tomando como base o trato vocal como um todo é possível imaginar que seja um tubo aberto ou fechado na ponta dependendo de qual fonema esteja sendo produzido. Sua pesquisa baseou-se fortemente na observação de imagens de raio-x de pessoas falando os mais diversos fonemas.

Diversos cientistas além de Fant buscaram modelar matematicamente a voz ou o aparelho vocal como um todo mas dada a complexidade do sistema os resultados nunca não foram totalmente satisfatórios. A modelagem física sofreu diversas especulações até o aparecimento de tecnologias como o raio-x ou a a endoscopia que permitissem observar as cavidades internas em movimento. Outra forte crítica à modelagem física é que muitos deles focam apenas no trato vocal mas abstraem

[1] https://www.youtube.com/watch?v=iwVu2BWLZqA

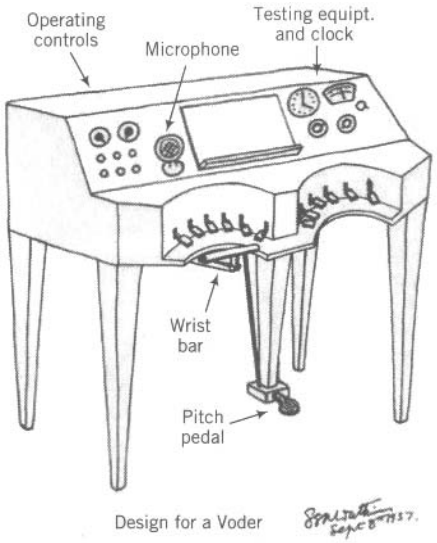
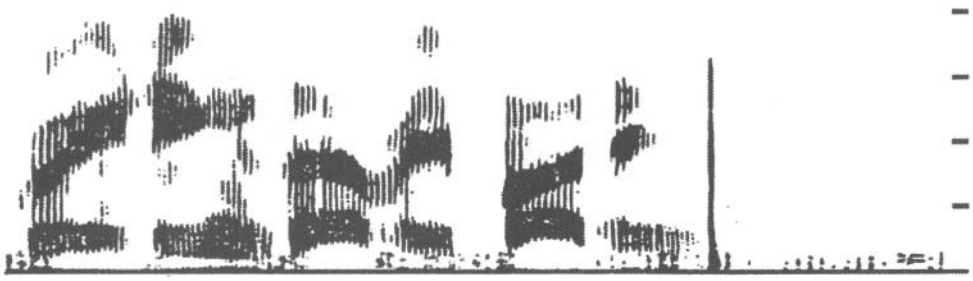
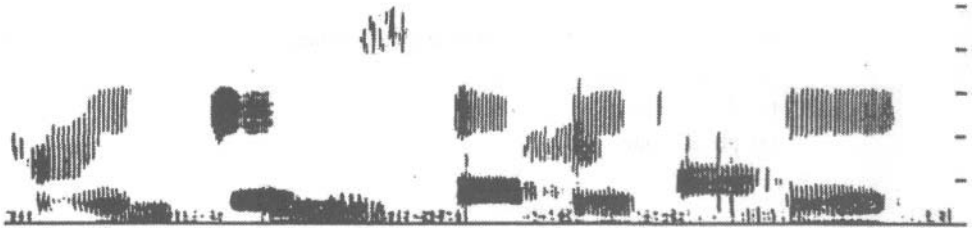


Figure 2.2: *Voder apresentador em 1939*



(a) *Humano*



(b) *Voder*

Figure 2.3: *Espectrograma da frase "greetings everybody"*

a importância de outras partes do corpo como a ressonância sofrida pela fala no crânio ou a amplitude da abertura do maxilar para produção da mesma.

2.2.4 Síntese Concatenativa

Em certo momento houve um interesse em gravar de alguma maneira instruções que pudesse ser replicada posteriormente para reprodução de um mesmo som. Tomando a mesma ideia dos cartões perfurados usados nos primeiros computadores criou-se uma pianola, que é um piano que lê um rolo responsável por apontar qual nota deve ser tocada e suas respectivas durações.

Os primeiros fonógrafos e gramofones datam do final do século XIX com um mecanismo de cone para coletar as vibrações sonoras registradas através de uma membrana e com uma sensível agulha registrar essas vibrações de modo que pudessem ser replicadas e gerar o som novamente. Por diversos fatores os primeiros aparelhos eram bastante limitados e não eram capazes de capturar nem todo o conteúdo da nossa fala nem toda a faixa de frequência que somos capazes de ouvir. Nossa faixa de frequência audível está entre 20 e 20kHz, aproximadamente enquanto a voz flutua entre 50 e 3400Hz em geral. Os primeiros gravadores, no entanto, capturavam apenas a faixa entre 250 e 2500Hz aproximadamente o que compreende a quase 70% da faixa representativa da voz.

A partir da evolução da gravação do som foi possível desenvolver a síntese concatenativa que, por ser originada completamente de um processo natural produzia resultados mais agradáveis ao ouvido do que os modelos matemáticos até então introduzidos.

Unidades Mínimas

Na síntese concatenativa gravam-se diversas *unidade mínimas* de um falante e as concatenamos para síntese posterior. Essas unidades mínimas variam de acordo com o sistema utilizado. Analisaremos primeiramente o padrão ocidental que nos é mais comum e então faremos alguns comentários sobre o padrão oriental com foco no japonês por ser uma língua com grande interesse na síntese de voz.

Usar cada letra como unidade mínima é trabalhoso e gera resultados imprecisos pois as letras apresentam diferentes comportamentos pareadas com umas e outras. O sistema de letra para fonema é frequentemente utilizado como suporte ao sistema de dicionário, que é o mais famoso. No sistema de dicionário uma grande quantidade de palavras tem seu(s) equivalente(s) fonético(s) representado(s) para consulta. Em geral a construção de um dicionário é trabalhosa por ter que traduzir tantas palavras de um formato para outro. Podemos citar dentre os dicionários o CMU

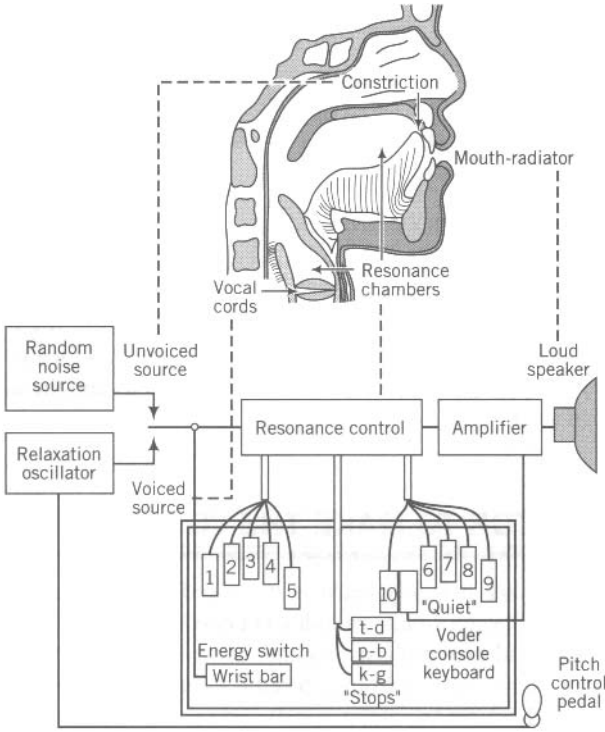


Figure 2.4: *Esquema de Filtragem de um Voder*

Dict, organizado pela Carnegie Mellon University (Uni17) e o dicionário fonético brasileiro Fala Brasil (NSKA08) criado pelo grupo da UFPA em 2008. Outra pesquisa importante focada na geração de um sistema de conversão grafema para fonema é o PETRUS (MZ14) mas esse sistema é baseado em HMM treinado a partir do mesmo dicionário produzido pelo grupo do Fala Brasil (Sil11).

Uma outra unidade mínima frequentemente citada na literatura é o fonema. O fonema é uma unidade mínima excelente pois existe uma quantidade finita e reduzida de fonemas mas muitos fonemas são influenciados pelos imediatamente anteriores ou posteriores de modo que a concatenação exclusivamente de fonemas acaba gerando aberrações e falhas prejudicando assim a síntese devido aos resultados artificiais. A síntese com unidades é de baixo custo computacional proporcional apenas ao esforço necessário apenas na junção das unidades. O custo computacional é contrastado porém com a grande massa de dados necessário para armazenar todas as unidades.

Uma das propostas para reduzir o problema de aberrações nas junções de fonemas é a gravação de difonemas (tradução livre de *diphone*) que são dois meio fonemas adjacentes. Pode ser entendido como a transição de fonemas e, desse modo, se torna um candidato que oferece estabilidade nas concatenações já que existe uma maior estabilidade na sustentação de um fonema do que na transição dos mesmos. A síntese a partir de difonemas ou apenas fonemas gera um banco de tamanho reduzido mas varia de tamanho de acordo com a linguagem apresentada (*e.g.* a quantidade de fonemas de Coreano difere de Inglês). Além disso é possível aplicar técnicas de DSP (***D**igital **S**ignal* **P**rocessing) para adicionar prosódia, movimento e entonação reduzindo a monotonia da voz. Técnicas como PSOLA, por exemplo, alteram o timbre da fala e podem ser até usadas para manipular a voz em uma certa melodia, como é utilizada em diversas montagens com políticos cantando músicas populares [2].

Nos idiomas orientais a própria formação da língua auxilia a construção de sintetizadores com unidades mínimas de cada linguagem. O japonês, por exemplo, é composto de diversas unidades mínimas com sons bem determinados em seus alfabetos fonéticos (Hiragana (Fig. 2.5) e Katakana). O alfabeto de Kanjis expressa uma ideia ou palavra e não auxilia em nosso estudo.

Tendo em vista que as unidades são muito bem definidas é compreensível que o japonês tenha

[2] https://www.youtube.com/user/baracksdubs

Figure 2.5: *Alfabeto de Hiragana com Fonético Ocidental Respectivo*

facilidade em adotar esse método de unidades naturalmente dada a estrutura natural da língua. Esse é um dos fatores mais importantes que devemos observar ao analisar o japonês como língua predominante nos idiomas de síntese com essa técnica.

A síntese com sentenças inteiras é a mais natural mas configura aplicações específicas demais para ser considerada como síntese e adota-se apenas o termo reprodução.

2.2.5 Síntese de Formantes

A síntese de formantes é interessante por não precisar de um falante para produzir amostras. Essa técnica se baseia em parâmetros que podem ser controlados como: frequência fundamental, níveis de voz e nasalamento do som no tempo. Através dessa parametrização esse sistema torna-se compacto e de baixo custo computacional mas traz com isso uma voz de sonoridade artificial. Mas ao mesmo tempo essa voz é mais facilmente inteligível em altas velocidades tornado-se uma vantagem para leitores de texto. Essa técnica é útil para sistemas com restrições de memória ou processamento. O controle total de todos os aspectos da síntese de formantes permite uma miríade de prosódias e entonações variando a expressão de sentimentos da voz sintetizada mas para isso requerem uma manutenção complexa de seus parâmetros no tempo.

2.2.6 Aplicações e Interesse Comercial

Atualmente diversos serviços cotidianos usam a síntese de fala com diferentes tipos de técnicas e unidades mínimas. Serviços de navegação muitas vezes misturam unidades mínimas de frases inteiras com instruções mais comuns e efetuam uma síntese concatenativa para leitura do nome de ruas.

A Yamaha atualmente vende o sistema de Vocaloids que é extremamente popular no japão. Sua voz mais popular é a da Vocaloid Hatsune Miku cuja popularidade extendeu-se de tal maneira que a voz ganhou um rosto, um modelo tridimensional e teve até apresentações em estádios em união ao artifício da projeção holográfica.

A síntese também é vendida como um serviço de terceiros, onde o interesse não é o modelo ou a técnica utilizada e sim métricas de negócio como volume de requisições, tempo de síntese, aplicabilidade no negócio, entre outros. Dos modelos comerciais que gabam dessas características e

efetivamente afirmam utilizar modelos de rede neural (por mais que não explicitem os modelos) são o Amazon Polly (Ama19) e o Google Cloud Text-to-Speech (Goo19). O Google utiliza efetivamente o serviço de síntese em outros de seus serviços como o Google Tradutor, Assistente e mais recentemente demonstrou o projeto Duplex que é capaz de interagir em todo o processo de compreensão, síntese e interação para uma dada atividade.

Avançando ao ferramental comercial interessado na síntese genérica de qualquer falante tivemos em 2016 a Adobe que exibiu na sua própria conferência um protótipo de aplicação chamada Adobe VoCo[3] com propriedades de edição similar a outros editores de áudio mas com a capacidade de sintetizar perfeitamente sentenças não faladas por um dado falante com uma quantidade mínima de áudio para treino e logo em seguida a fundação da startup canadense lyrebird[4] fundada por estudantes da MILA (University of Montreal - Montreal Institute for Learning Algorithms) pupilos do Yoshua Bengio, um dos patronos de vários artigos importantes na área de redes neurais. A simples sugestão de tais aplicações gerou uma grande preocupação quanto à possíveis má utilizações e a falta de assinaturas ou métodos que permitissem a verificação de tais sínteses. Enquanto não desenvolveu-se uma solução para esse problema houve continuidade de pesquisa apenas no meio acadêmico.

Cronologicamente temos a constante participação de grandes empresas na continuidade da pesquisa pela síntese de fala genérica como podemos ver na organização de trabalhos abaixo:

- Em Setembro de 2016 o **Google** publicou o primeiro resultado do WaveNet (vdODZ+16) que seria o primeiro modelo neural responsável pela síntese de áudio;
- Em fevereiro de 2017 o **Baidu** publicou resultados do Deep Voice (ACC+17) que seria o primeiro modelo completamente composto por redes neurais a atingir a síntese de fala;
- Em março de 2017 o **Google** publicou a primeira versão de seu modelo de síntese de fala ponta a ponta, o Tacotron (WSRS+17)
- Em maio de 2017 **Baidu** novamente publicou resultados do seu modelo Deep Voice melhorado nomeando-o Deep Voice 2 (ADG+17). Houve melhorias de performance no treinamento devido a um kernel implementado diretamente nas GPU contornando assim os gargalos presentes nas trocas de informação com os frameworks anteriormente utilizados;
- Em julho de 2017 o **Facebook** publicou o VoiceLoop (**?**)[5], o foco desse trabalho foi a síntese a partir de exemplos naturais, ou seja, sem se preocupar com a correlação de sinal-ruído;
- Em outubro de 2017 **Baidu** publicou a versão final de seu modelo Deep Voice, o Deep Voice 3. Esse modelo usou um sistema mais simples que os anteriores o que fez com que seu treino e inferência fossem mais rápidos e a convergência do modelo pudesse ser obtida mais rapidamente já que haviam menos parâmetros a otimizar. Esse modelo também se utilizou largamente de convoluções permitindo uma alta paralelização;
- Em dezembro de 2017 o **Google** publicou o Tacotron 2(**?**), uma versão incrementada do modelo apresentado no início do mesmo ano.
- Em Fevereiro de 2018 o **Baidu** (**?**) reavaliou seus modelos anteriormente publicados para confrontar o modelo do Facebook nesse trabalho focou em desenvolver uma solução de readaptação dos modelos anteriores

Todos esses modelos compartilham das mesmas ideias construídas e dissecadas nos modelos de síntese estudados anteriormente. Gostaria de usar nesse trabalho uma implementação do modelo do Deep Voice cujo modelo básico pode ser visto na Fig. 2.6. O intuito desse trabalho é observar

[3] https://theblog.adobe.com/lets-get-experimental-behind-the-adobe-max-sneaks/
[4] https://lyrebird.ai/
[5] https://research.fb.com/publications/voiceloop-voice-fitting-and-synthesis-via-a-phonolgoical-loop/

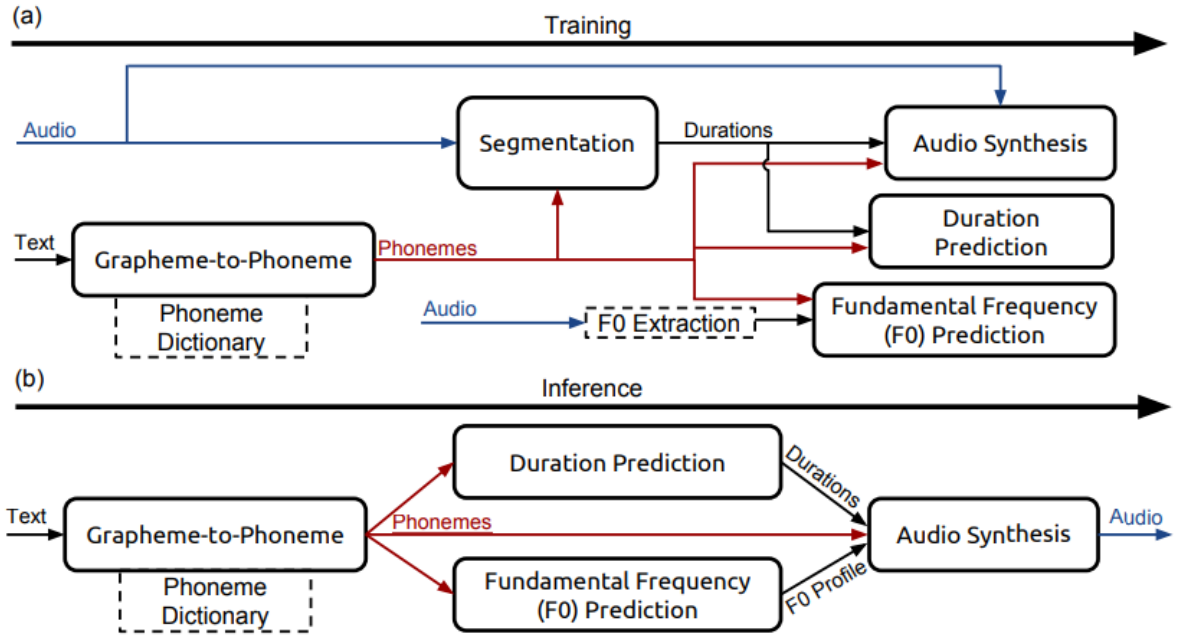


Figure 2.6: *Blocos de Síntese do Deep Voice*

especificamente a alteração do idioma de entrada que confere um impacto maior na rede de entrada, no caso, o modelo grafema para fonema que tem suas entradas alinhadas ao modelo de segmentação. Esse modelo tem como responsabilidade fazer a conversão de todo texto de entrada para algum modelo fonético (como o IPA ou o SAMPA) assim como alinhar esses símbolos escolhidos para alguma sequência de áudio de entrada. O modelo treinado nesses casos é um modelo de atenção que será detalhado na seção 2.4. O modelo responsável pela síntese é o WaveNet. O modelo de extração de F0 da voz é treinado com uma outro algoritmo de extração de frequência fundamental para cada segmento determinado.

2.3 Alfabetos Fonéticos

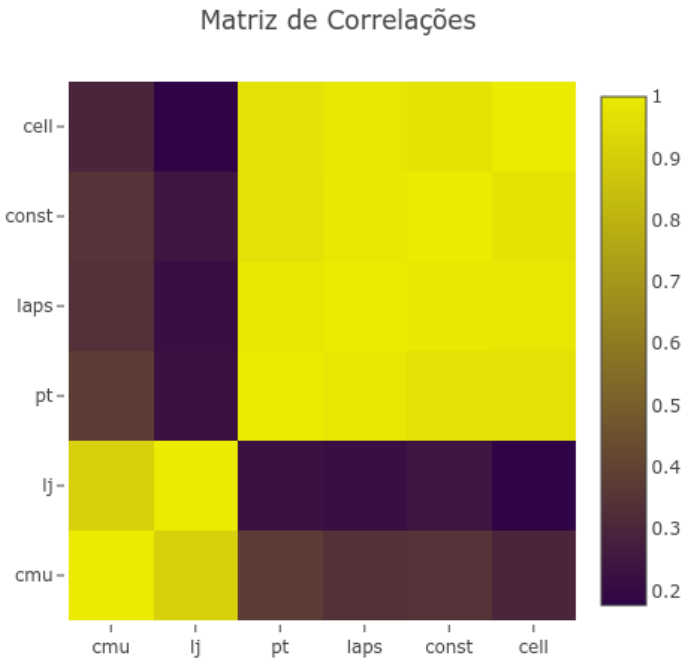
A quantidade de fonemas possíveis de serem expressados pelo trato vocal humano é finito. Não é comum uma linguagem se utilizar de todos os fonemas possíveis na fala humana e a presença de uma quantidade maior ou menor de fonemas também acaba determinando a facilidade de se aprendê-la.

Para garantir uma internacionalização linguisitas criaram o IPA (*International Phonetic Alphabet*) que pode ser consultado na tabela no Apêndice A. Esse alfabeto vem passando por constantes revisões desde sua criação inicial em 1886. Devido a presença de diversos símbolos fora do padrão latino houve a necessidade da criação de um outro alfabeto que servisse de conversão intermediária e pudesse ser facilmente lido por computadores (anterior à implementação do padrão UTF-8). Para a fácil interpretação dos computadores foi criado o alfabeto SAMPA que consiste na mesma ideia de representar a fonética humana com símbolos mas o foco agora foi a fácil implementação nos computadores.

Um dos maiores dicionários fonéticos em inglês, o CMU Dict(Uni17), se utiliza desse padrão SAMPA. Através da análise de um dicionário fonético temos como observar a frequência de fonemas relativos em uma linguagem e assim estabelecer uma comparação a outra linguagem.

Uma comparação apenas por dicionário pode não ter impacto no balanço de fonemas utilizados efetivamente na fala, assim observamos a frequência relativa dos fonemas efetivamente utilizados na análise abaixo

LJSPeech

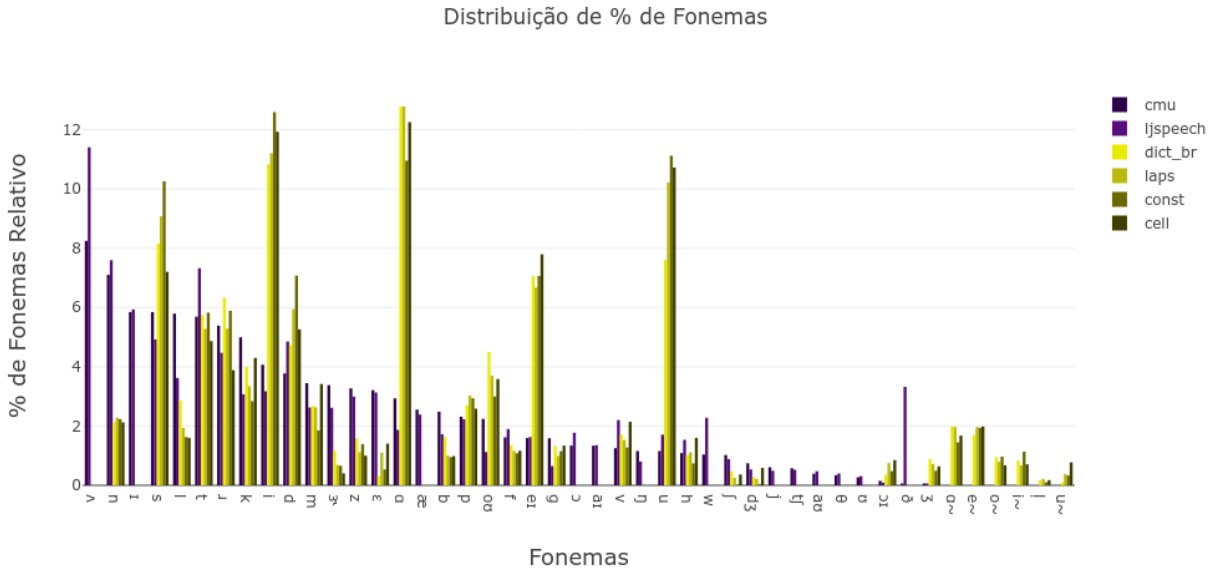


(a) *Mapa de Temperatura*

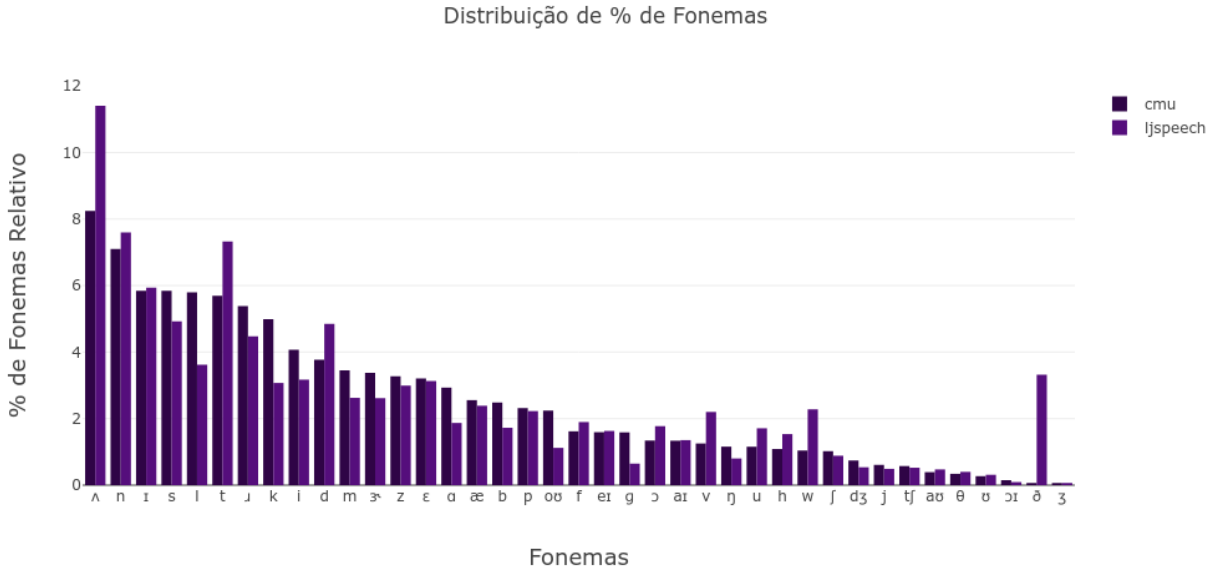
	cmu	lj	pt	laps	const	cell
cmu	1.00	0.91	0.38	0.34	0.35	0.29
lj	0.91	1.00	0.23	0.22	0.24	0.18
pt	0.38	0.23	1.00	0.98	0.96	0.97
laps	0.34	0.22	0.98	1.00	0.99	0.99
const	0.35	0.24	0.96	0.99	1.00	0.97
cell	0.29	0.18	0.97	0.99	0.97	1.00

(b) *Matriz de Correlação*

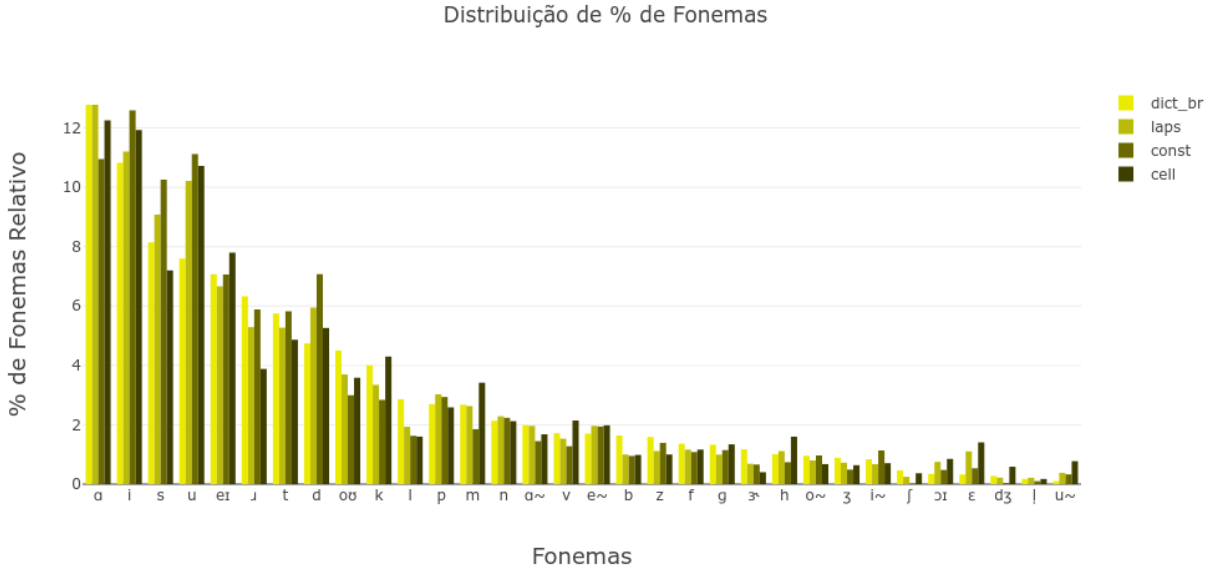
Figure 2.7: *Observando a matriz de correlação observamos alta correlação nos fonemas de uma mesma língua nos conjuntos de dados.*



(a) *Participação Relativa em Todas as Línguas*



(b) *Participação Relativa no Inglês*



(c) *Participação Relativa no Português*

Figure 2.8: *Distribuição de Fonemas por conjunto de dados*

Para a análise tomamos dois dicionários fonéticos, um em inglês e um em português, e alguns conjuntos de falas transcritas que utilizaremos no processo de treino. Usamos como base para o inglês o LJSpeech e para o português três conjuntos de fala: o LAPS, a Constituição, ambos fornecidos pelo projeto FalaBrasil, e o conjunto de fala extraído de vídeos do youtube, conforme detalhado na seção 3.1.1.

Para montar a análise entre línguas utilizamos um método de conversão apontado pela tabela de fonemas internacional, disponível no Apêndice A. Dessa forma garantimos o alinhamentos dos fonemas corretos entre idiomas. Para a análise dos trechos falados fomos restritos ao alinhamentos das transcrições disponíveis com os respectivos dicionários de cada língua. Naturalmente esse foi um processo com perdas conforme apontado abaixo:

1. LJSpeech - Perda de 01139 palavras, correspondente a 08.28%
2. LAPS - Perda de 00046 palavras, correspondente a 01.68%
3. Constituição - Perda de 00526 palavras, correspondente a 09.87%
4. YouTube - Perda de 03344 palavras, correspondente a 50.08%

Podemos perceber que muitos dos termos obtidos dos dados do youtube não foram alinhados com algum termo do dicionário. Isso tem raiz no grande uso de neologismos, aumentativos, diminutivos, estrangeirismos e outras formas de variação da linguagem formal.

2.4 Redes Neurais e Aprendizado de Representações

O tópico redes neurais gera grande confusão naqueles que são novos no campo por englobar muitas técnicas com diferentes finalidades. O intuito desse trabalho não é dissertar sobre todos os aspectos que compõem o campo de Aprendizado Profundo (*Deep Learning*), Redes Neurais ou do conceito geral de Aprendizado de Representações. Apresentaremos nessa seção um breve resumo dos blocos importantes à compreensão deste trabalho pontuados por referências onde o leitor pode encontrar um detalhamento maior dos mesmos além de diversos outros tópicos relacionados nesse conjunto de técnicas.

Este trabalho pensa na sequência de letras, fonemas, palavras e sons como uma cadeia temporal sequencial de informações. A literatura de técnicas para cadeias temporais tem origem no interesse de se trabalhar com texto. Como o próprio trabalho se propõe a receber uma entrada de texto damos a intuição das técnicas também com texto.

2.4.1 Representações Densas

Um dos problemas básicos para tratar palavras em aprendizado de máquina é a dificuldade de se obter um padrão das mesmas como entrada para os algoritmos. Quando lidamos com texto usualmente não temos um tamanho de sequência bem definidas. No contexto de palavras temos palavras com diferentes tamanhos e no contexto de frases temos sentenças com diferentes comprimentos.

Uma das primeiras técnicas usadas foi a construção de um vetor de contagem de palavras. Nessa técnica cada sentença é abstraída em um vetor esparso com comprimento do vocabulário do corpus. Essa técnica mitiga o problema de falta de padrão pois todas as palavras teriam alguma representação vetorial com uma mesma dimensão. O problema dessa técnica além da dificuldade de se lidar com vetores esparsos é a completa ortogonalidade entre todos os termos, isso é, assume-se que todas as palavras seriam completamente independente umas das outras. Essa premissa é claramente falsa quando pensamos na morfologia e sintaxe da nossa língua. Isso também acontece quando observamos os fonemas que são divididos em classes de acordo com o som que produzem e da maneira que são produzidos.

Palavras que compartilhem uma mesma característica morfológica ou que possuam proximidade semântica poderiam estar agrupadas em um mesmo grupo de contextualização. Baseado nessa ideia

podemos pensar que em um corpus grande suficiente teríamos uma quantidade razoável de relações semânticas para ser aprendidas. Essa é a intuição do word2vec (MSC+13) onde temos duas técnicas de aprendizado: a CBOW (Cumulative Bag of Words) e a SkipGram. Na CBOW várias palavras são alimentadas e buscamos prever uma única palavra de contexto enquanto na SkipGram a palavra de contexto é alimentada e buscamos prever as palavras dentro daquela janela. O intuito dessa técnica é exatamente tentar capturar a semântica das palavras baseado nas palavras que aparecem próximas a ela. Esse tipo de semântica só pode ser bem generalizado em corpus grandes tendo em vista as mais diversas situações que uma mesma palavra poderia se relacionar com outras. A utilização de corpus pequenos tende a gerar modelos pouco genéricos que acabam englobando apenas um pequeno domínio.

Outras técnicas incrementaram pontos fracos do word2vec como a Glove (PSM14) ou fastext (**?**). Essas técnicas são comumente usadas para capturar relações entre palavras mas podem ser usadas em outros níveis também, com a nível de caracter, sentença, parágrafo, entre outros. No caso do nosso problema esse módulo é responsável por capturar uma representação própria para a palavra visando sua representação fonética. Para agregar essa informação precisamos de outra célula, a célula recorrente.

2.4.2 RNN - Redes Neurais Recorrentes

As células recorrentes foram pensadas para capturar informações em uma dada cadeia de dados. Elas funcionam como os neurônios simples clássicos mas uma das entradas que recebem é também a saída da iteração do passo anterior (Fig. 2.9).

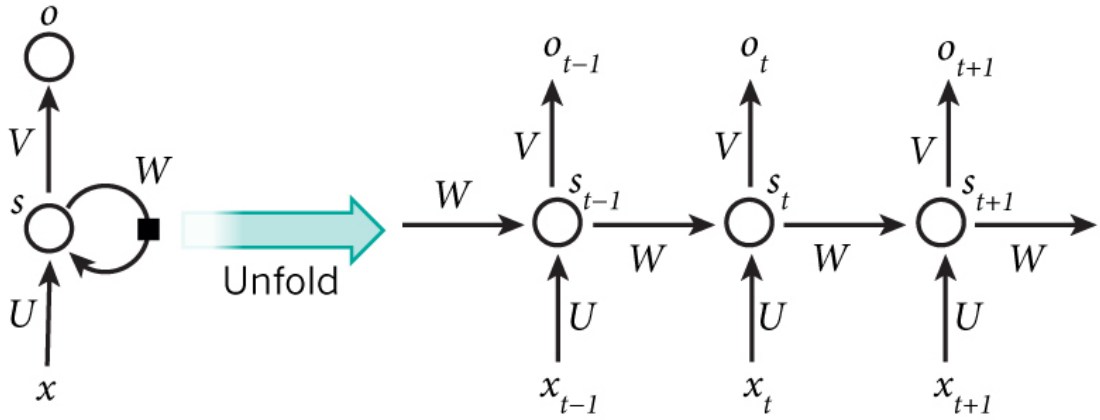


Figure 2.9: *Célula Recorrente*

No entanto essas células inserem um desafio computacional quanto à complexidade de sua propagação de erros. Comumente nos neurônios simples existe uma função bem definida cujo comportamento pode ser ajustado através do algoritmo de propagação retroativa de erros (*backpropagation*) mas ao adicionarmos uma dimensão extra (tendo em vista que é uma sequência) a propagação sofre um problema. O primeiro problema fácil de perceber é que a medida que as sequências ficam mais longas a propagação demora cada vez mais para atingir o início da sequência dada a necessidade de se recalcular os gradientes com os erros para cada passo anterior do tempo na sequência. O segundo problema provém do mesmo problema onde caso a o módulo da matriz de pesos seja maior que 1 o gradiente explode à medida que é propagado e não somos capazes de aprender nada (*exploding gradient*) caso o módulo seja menor que 1 o problema oposto ocorre onde o erro propagado rapidamente tende a 0 e ficamos incapazes de propagar erros mais que alguns passos no passado (*vanishing gradient*). O comportamento do gradiente pode ser controlado a partir da matriz de inicialização dos pesos no passo inicial. Desde que a inicialização seja bem calculada com valores que evitem a explosão e a inércia dos valores. (IGC16)

Numa célula RNN tradicional toda informação é passada para a célula seguinte. Isso torna

sequências muito longas ou cujas informações relevantes sejam necessárias vários passos depois muito difíceis de codificar. Uma tentativa de se controlar a quantidade de informação passada em cada passo foram as unidades GRU (*Gated Recurrent Unit*) e LSTM (Long Short Term Memory). Essas duas unidades (Fig 2.10) possuem sistemas de passagem parcial de informações, seus portões, cujos parâmetros também são aprendidos durante o treino. A medida que acrescentamos parâmetros numa célula geramos uma crescente sobrecarga de informações a serem computadas em cada passo de modo que essa técnica possui um custo computacional no treinamento do modelo.

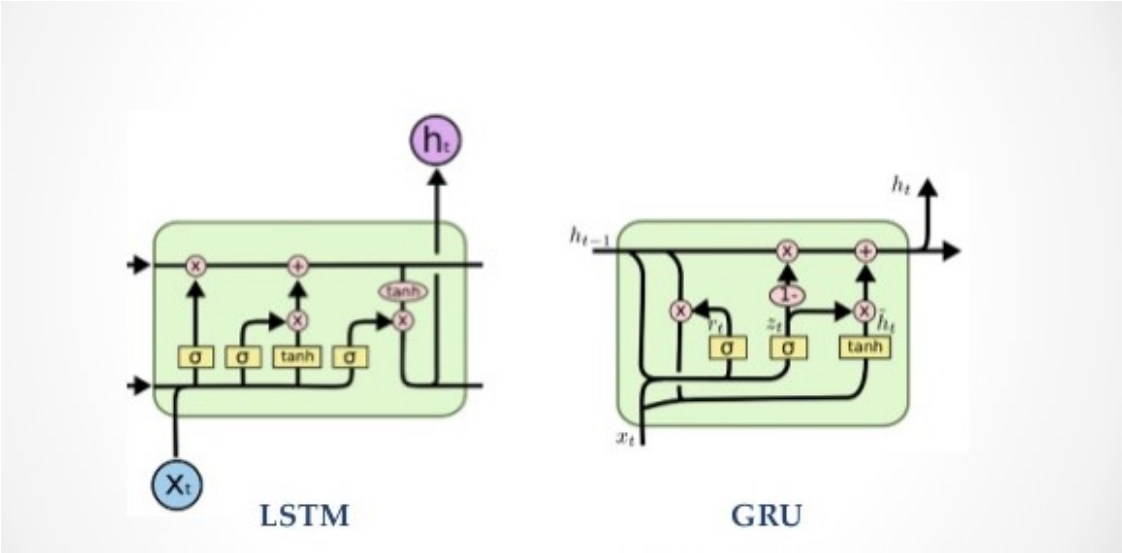


Figure 2.10: *Células Recorrentes*

2.4.3 Conectist Temporal Classification (CTC) e Beam Search

Em um modelo encoder-decoder o decoder pode ter algumas estratégias para executar o processo de decodificação. A mais simples é predizer apenas um valor (seja ele uma classe ou um valor regredido). Essa abordagem é bastante agressiva pois qualquer classificação errônea seria propagada pelo modelo de decodificação. Uma estratégia pensada para contornar esse problema é a predição das melhores classes que se adéquem a algum limiar. Desse modo cada passo pode predizer várias classes e a saída fica sendo a sequência de probabilidades das classes a serem preditas. Ao final da predição da cadeia analisa-se essa sequência e montamos a saída através de uma programação dinâmica onde busca-se maximizar as probabilidades entre as unidades.

Tomando um exemplo de tradução por exemplo podemos ter o seguinte exemplo: o encoder codifica a palavra em português "gato" para um vetor intermediário denso; o decoder com CTC pode decodificar esse vetor de várias maneiras como podemos ver abaixo, onde o espaço de quebra de letra é o $< b >$:

- $ccc < b > a < b > ttttt < b >$
- $c < b > aaa < b > tt < b >$
- $ccc < b > aaoaa < b > t < b >$

Através do algoritmo de programação dinâmica Beam Search buscamos maximizar a probabilidade da predição da cadeia. O algoritmo busca maximizar a probabilidade dentre as unidades previstas de se alinharem entre um mesmo bloco e entre blocos. Cada bloco é o conjunto de predições que não foram quebradas por um caracter de quebra ($< b >$). Assim também conseguimos propagar os gradientes na rede e somos capazes de mapear todas as sequências para a palavra "cat". Mesmo classificações incorretas em algum tempo podem ser corrigidas pela maximização de probabilidade de cada letra, no caso.

As unidades mínimas típicas de texto são letras e palavras. O alinhamento do encoder-decoder num sistema de tradução neural é então de letra para letra ou de palavra para palavra. No nosso modelo grafema-fonema a entrada do modelo é texto e a saída é fonema. O alinhamento ocorre com os trechos de palavra conhecidos com algum fonema previsto na saída. O mesmo princípio explicado acima com letras se verifica e os fonemas são alinhados a uma saída seccionando a entrada paralela a um trecho respectivo de texto e som.

2.4.4 Convoluções Autoregressivas Dilatadas

O modelo de síntese mais popular atualmente é o WaveNet (vdODZ+16). O intuito desse módulo é oferecer uma alternativa à concatenação de unidades de síntese mínimas utilizadas. Na concatenação tradicional ocorrem aberrações fruto da natureza seccionada das unidades mínimas.

No modelo do WaveNet a geração de amostras de áudio é probabilística e atrelada as amostras anteriores. Comparativamente pode-se pensar que esse modelo utiliza uma unidade mínima como a unidade do áudio amostrado sem tratamento. Podemos perceber que isso trás um problema similar ao do gradiente nos modelos recorrentes pois as sequências de áudio com alta taxa de amostragem acabam sendo muito longas mesmo para poucos segundos de áudio. Para contornar esse problema o trabalho propôs duas técnicas combinadas: a utilização de uma janela onde apenas alguns passos anteriores seriam considerados na síntese de uma nova amostra; e a utilização das próprias saídas nas entradas de modo a eliminar as possíveis anormalidades presentes na transição de unidades.

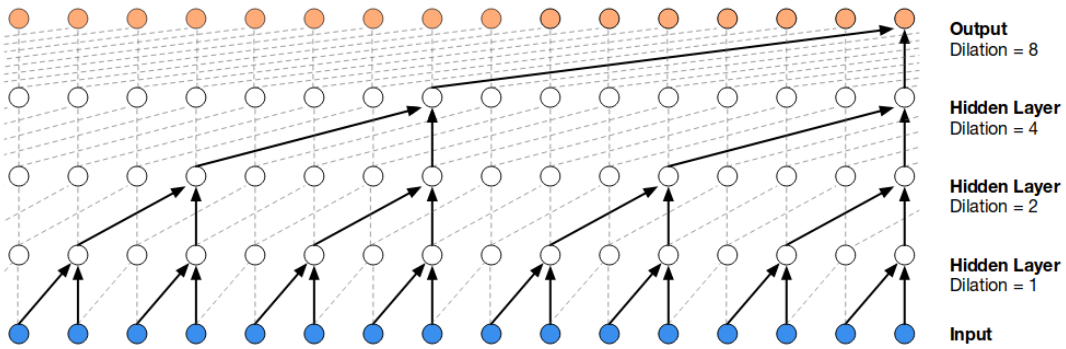


Figure 2.11: *Convolução Dilatada*

A primeira técnica efetua convoluções dilatadas (Fig 2.11) onde a convolução é aplicada com espaçamentos de acordo com o valor alocado. O caso especial onde a dilatação é 1 configura a convolução tradicional.

A segunda técnica consiste numa estratégia que remete aos modelos de recorrência pois utiliza cada saída como entrada para as convoluções futuras. Essa estratégia pode ser executada rapidamente no treino do modelo pois as entradas são facilmente paralelizáveis e o modelo pode efetuar diversas operações de convolução simultaneamente. Entretanto quando essa estratégia é utilizada para predição é necessário esperar a geração de cada amostra para computar as amostras posteriores removendo o fator paralelizável e tornando-se um gargalo para a síntese em tempo real. Propostas posteriores (ACC+17**?**) buscaram outras estratégias para alcançar performances melhores de tempo de execução do modelo

Chapter 3

Metodologia

O intuito na seleção de áudios bons tem os mesmos princípios do problema de ASR (*Audio Speech Recognition*). Reduzir a relação sinal-ruído o máximo possível, tomar amostras com a menor variabilidade de entre falantes possível (e.g. Sotaques, idades, gênero, comprimento do trato vocal), tomar amostras com a melhor qualidade de áudio possível reduzindo qualquer efeito proveniente do ambiente (e.g. Reverberação), tomar amostras sem interseção entre falantes (e.g. *Cocktail Party Problem*).

Começamos esse capítulo apresentando conjuntos de dados públicos e discutindo métodos possíveis para a construção de um conjunto de dados próprio com dados de outros falantes. Após a discussão dos conjuntos de dados disponíveis e da seleção de um método para construção de um conjunto próprio analisamos os dados obtidos prevendo possíveis comportamentos provenientes do modelo final. Retomamos o modelo introduzido anteriormente e apontamos as mudanças necessárias para sua execução na língua portuguesa. Detalhamos então o processo de treino que pôde ser mais facilmente explorado e a partir dos melhores modelos selecionamos alguns resultados preliminares para discussão.

3.1 Construção do Dataset

Para trabalhar o treinamento da mesma rede na língua portuguesa buscamos encontrar primeiro conjuntos de dados já previamente construídos e disponíveis em domínio público. Alguns dos conjuntos de dados encontrados, no entanto não englobavam as necessidades mínimas de treino do modelo. O conjunto de dados[1] exibe apenas segmentos fonéticos gravados por um falante. Nesse conjunto de dados o problema reside na impossibilidade de se extrair a transição entre diferentes fonemas e no desbalanceamento de fonemas comparativamente a uma situação real.

O conjunto de dados [2] possui frases de vários falantes e com várias frases distintas. Os inputs foram gravados manualmente pelos usuários em diversas fontes diferentes estando suscetíveis à ruídos externos e inconsistências de taxas de amostragem já que cada usuário usou seus próprios meios de gravação. Cada falante decide um conjunto de frases que gravaria de modo que temos vários falantes com poucas frases.

Por fim o conjunto de dados disponível em [3]

Os conjuntos de dados citados anteriormente possuem a facilidade de estar disponíveis para download mas dadas as restrições de falantes, extensão de frases naturais faladas e diversos ambientes estabelecemos possíveis técnicas genéricas para facilitar uma possível construção posterior por outros usuários com outros falantes.

Dado um arquivo de áudio contendo fala e sua transcrição correspondente computar um alinhamento forçado consiste em, para cada fragmento da transcrição, determinar o intervalo respectivo no trecho de áudio contendo o respectivo fragmento. Os fragmentos podem ter diversas

[1] https://www.kaggle.com/jonascarvalho/brazilian-portuguese-phonemes-audio/version/2
[2] http://www.voxforge.org/pt/downloads
[3] http://labvis.ufpa.br/falabrasil/downloads/

granularidades como as já discutidas anteriormente (fonema, palavra, sentença, parágrafo). O caso base, onde a granularidade é a maior possível alinha somente o texto inteiro ao áudio inteiro, não frequentemente útil. Dentre os exemplo práticos podemos destacar a construção de arquivos de legenda para sincronização, .

Um alinhamento forçado é um dado alinhamento de um áudio de entrada e um texto de entrada.

Comumente utilizado para alinhamento de legendas quando o texto já está disponível mas o alinhamento ainda não foi feito manualmente. Existem diversas ferramentas reunidas que já executam essa atividade [4] cuja maioria se destaca pela presença de HMMs como técnica mais utilizada. Num momento inicial experimentamos a construção de um conjunto de dados teste com a ferramenta aeneas [5] mas os resultados tanto em performance quanto em qualidade empírica não foram satisfatórios.

O alinhamento teste demorou até três vezes o tempo de comprimento do próprio áudio para gerar uma saída. A saída muitas vezes sofria com um pequeno desalinhamento no início do texto que descarrilhava uma falha sequencial em cascata.

Nas próximas seções discutimos as estratégias estudadas para alinhamento de áudios.

3.1.1 Álinhamento de Legendas do YouTube

Uma outra alternativa que surgiu foi utilizar os áudios e legendas do YouTube que já tenham sido avaliadas manualmente. O Youtube atualmente possui dois esquemas de legenda: em um a legenda é gerada automaticamente a partir do áudio do vídeo e fica disponível através da opção Legendas geradas automaticamente; essa legenda também é disponibilizada à quem fez upload do vídeo para avaliação e possível correção manual, tornando-se posteriormente disponível como uma legenda da linguagem definida.

Selecionamos um conjunto de vídeos com expectativa de reduzir o ruído de fundo e possíveis alterações no áudio que pudessem atrapalhar a generalização do modelo. Nessa expectativa o conjunto de vídeos selecionados para teste foi do estilo vlog onde o falante discorre num monólogo com a câmera. Fizemos o download dos dados com um script python com auxílio da biblioteca youtube-dl que possibilita o download apenas do áudio dos vídeos assim como a legenda dos mesmos. Os áudios e legendas baixados apresentaram qualidade satisfatória e foram posteriormente seccionados em arquivos menores para facilitar a compreensão do conteúdo pelos trechos sequencias do modelo que, como citado anteriormente, sofre com a captura de informações em sequências “muito longas”

[4] https://github.com/pettarin/forced-alignment-tools
[5] https://www.readbeyond.it/aeneas/

Chapter 4

Cronograma

4.1 Cronograma Previsto

16/02 Incrementar conceitos no aspecto técnico do aprendizado de máquina, do processamento de sinais, dos termos de fonoaudiologia e sua intercessão com computação. Escrever metodologia utilizada na construção dos dados, na alteração proposta pra rede e no treinamento do modelo

23/02 Selecionar os melhores modelos treinados e gerar frases sintéticas para análise

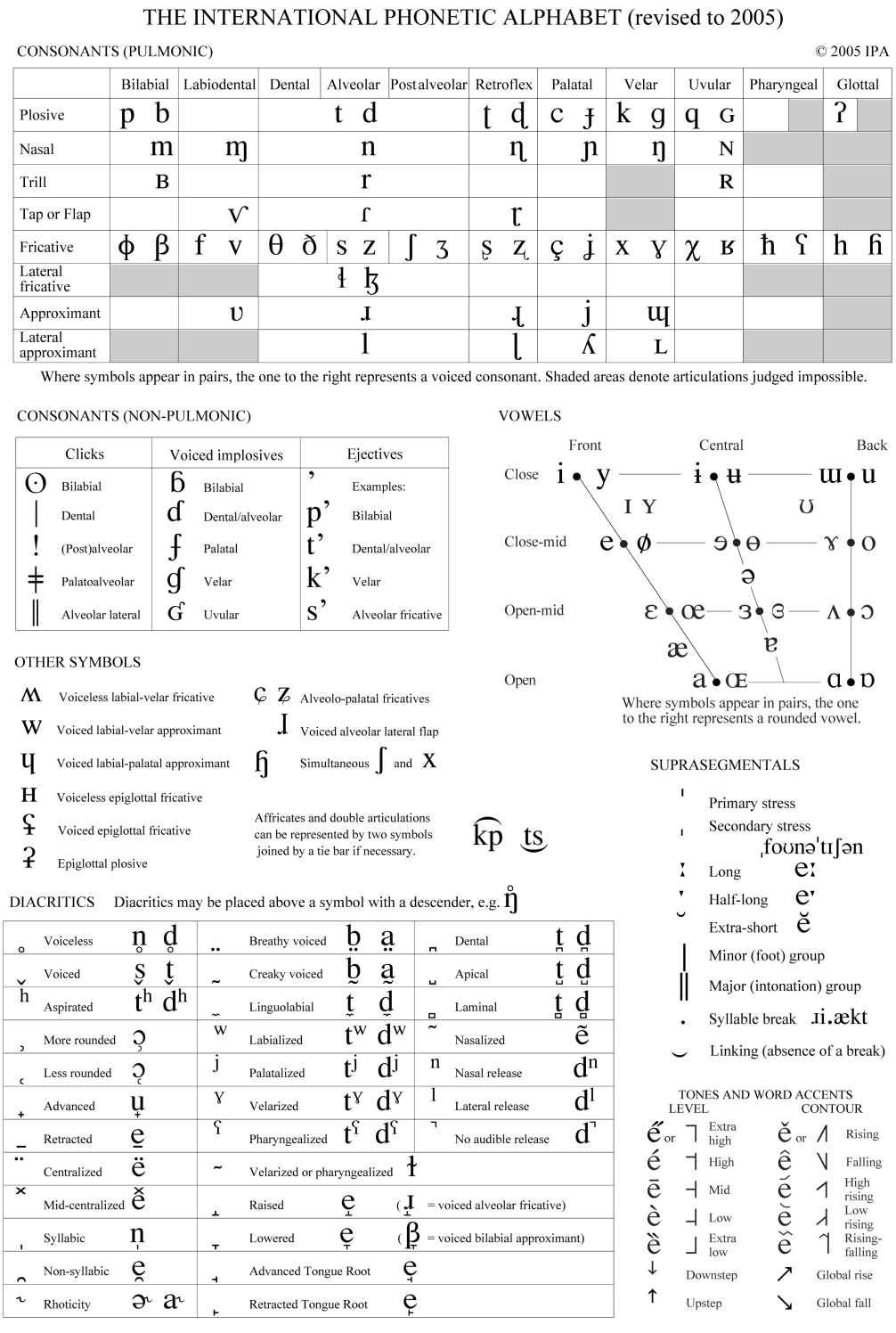
02/03 Finalizar teste e estabelecer as devidas análises estatísticas comparativamente em relação ao MOS dos trabalhos já publicados

09/03 Estabelecer especificações de hiperparametrização das melhores redes e respectivos resultados obtidos no trabalho escrito

16/03 Verificar reprodutibilidade do código e do experimento. Solicitação de Fork no repositório do Github

Appendix A

A.1 Tabela IPA



Bibliography

[ACC+17] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Greg Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta e Mohammad Shoeybi. Deep voice: Real-time neural text-to-speech. *CoRR*, abs/1702.07825, 2017. 11, 18

[ACP+18] Sercan Ömer Arik, Jitong Chen, Kainan Peng, Wei Ping e Yanqi Zhou. Neural voice cloning with a few samples. *CoRR*, abs/1802.06006, 2018.

[ADG+17] Sercan Ömer Arik, Gregory F. Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman e Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *CoRR*, abs/1705.08947, 2017. 11

[Ama19] Amazon. Amazon polly. https://aws.amazon.com/polly/, jan 2019. 11

[CDG+18] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew P. Aylett, João P. Cabral, Cosmin Munteanu e Benjamin R. Cowan. The state of speech in HCI: trends, themes and challenges. *CoRR*, abs/1810.06828, 2018.

[DSSO18] Tijana Delić, Sinisa Suzic, Milan Sečujski e Vladimir Ostojic. Deep neural network speech synthesis based on adaptation to amateur speech data. *IcETRAN 2018*, 06 2018.

[DT50] Homer Dudley e T. H. Tarnoczy. The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, 22(2):151–166, 1950. 6

[Dud55] Homer Dudley. Fundamentals of speech synthesis. *Journal of Audio Engineering Society*, 3:170–185, 1955.

[Fan60] Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960. 7

[fla94] *Voice Communication Between Humans and Machines*. National Academies Press, jan 1994.

[GEB15] Leon A. Gatys, Alexander S. Ecker e Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[GME11] Ben Gold, Nelson Morgan e Dan Ellis. *Synthetic Audio: A Brief History*, páginas 9–20. John Wiley & Sons, Inc., 2011. 6

[Goo] Google. Cloud text-to-speech. https://cloud.google.com/text-to-speech/. acessado em 03/2019.

[Goo19] Google. Cloud text-to-speech - speech synthesis. https://cloud.google.com/text-to-speech/, jan 2019. 11

[IGC16] Yoshua Bengio Ian Goodfellow e Aaron Courville. The challenge of long-term dependencies. Em *Deep Learning*, chapter 10, páginas 401–404. MIT Press, 2016. http://www.deeplearningbook.org. 16

[JB11] Emanuël A.P. Habets (auth.) Jacob Benesty, Jingdong Chen. *Speech Enhancement in the STFT Domain*. Springer, 2011.

[JGB+16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou e Tomas Mikolov. Fasttext.zip: Compressing text classification models, 2016. cite arxiv:1612.03651Comment: Submitted to ICLR 2017.

[JMD+17] Zeyu Jin, Gautham J. Mysore, Stephen Diverdi, Jingwan Lu e Adam Finkelstein. Voco. *ACM Transactions on Graphics*, 36(4):1–13, 2017.

[JRDJ93] John G. Proakis John R. Deller Jr., John H. L. Hansen. *Discrete-Time Processing of Speech Signals (IEEE Press Classic Reissue)*. Wiley, 1993.

[JZW+18] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno e Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *CoRR*, abs/1806.04558, 2018.

[KN18] Souvik Kundu e Hwee Tou Ng. A question-focused multi-factor attention network for question answering. *CoRR*, abs/1801.08290, 2018.

[Mas07] Dominic W. Massaro. *What Ar e Musical Paradox and Illusion?* American Journal of Psychology, 2007. 2

[Mic] Microsoft. Text to speech. https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/. acessado em 03/2019.

[MKXS18] Bryan McCann, Nitish Shirish Keskar, Caiming Xiong e Richard Socher. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730, 2018.

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado e Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. 16

[MZ14] Bokan A. Marquiafavel, V. e C. Zavaglia. *PETRUS: A rule-based grapheme-to-phone converter for Brazilian Portuguese*. Tese de Doutorado, COPPE UFRJ, 2014. 9

[MZSX18] Sewon Min, Victor Zhong, Richard Socher e Caiming Xiong. Efficient and robust question answering from minimal context over documents. *CoRR*, abs/1805.08092, 2018.

[NSKA08] Nelson Neto, Patrick Silva, Aldebaro Klautau e Andre Adami. Spoltech and ogi-22 baseline systems for speech recognition in brazilian portuguese. *Lecture Notes in Computer Science Computational Processing of the Portuguese Language*, página 256–259, 2008. 9

[PPG+17] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman e John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *CoRR*, abs/1710.07654, 2017.

[PSM14] Jeffrey Pennington, Richard Socher e Christopher D. Manning. Glove: Global vectors for word representation. Em *Empirical Methods in Natural Language Processing (EMNLP)*, páginas 1532–1543, 2014. 16

[Sil11] D. Silva. *Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em HMM.* Tese de Doutorado, Programa de Pós-Graduação em Engenharia Elétrica, COPPE, Universidade Federal do Rio de Janeiro, RJ, 2011. 9

[SPW+17] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis e Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017.

[TT12] Manjul Tiwari e Maneesha Tiwari. Voice - how humans communicate? *Journal of Natural Science, Biology and Medicine*, 3(1):3, 2012.

[TTH18] Yi Tay, Luu Anh Tuan e Siu Cheung Hui. Multi-pointer co-attention networks for recommendation. *CoRR*, abs/1801.09251, 2018.

[Tun05] Volkan Tunalı. *A Speaker Dependent, Large Vocabulary, Isolated Word Speech Recognition System for Turkish.* Tese de Doutorado, 07 2005. 2

[TWPN17] Yaniv Taigman, Lior Wolf, Adam Polyak e Eliya Nachmani. Voice synthesis for in-the-wild speakers via a phonological loop. *CoRR*, abs/1707.06588, 2017.

[Uni17] Carnegie Mellon University. The cmu pronouncing dictionary, 2017 (acessado em Julho de 2017). http://www.speech.cs.cmu.edu/cgi-bin/cmudict. 9, 12

[vdODZ+16] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior e Koray Kavukcuoglu. Wavenet: A generative model for raw audio. Em *Arxiv*, 2016. 11, 18

[vdOLB+17] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C. Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov e Demis Hassabis. Parallel wavenet: Fast high-fidelity speech synthesis. *CoRR*, abs/1711.10433, 2017.

[VK91] W Von Kempelen. *Le Mechanisme de lapavola, suivi de la Description d'une machine parlante.* J.V. Degen, 1791. 1, 6

[WJ16] Shuohang Wang e Jing Jiang. Machine comprehension using match-lstm and answer pointer. *CoRR*, abs/1608.07905, 2016.

[WSRS+17] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark e Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. *CoRR*, 2017. 11

[XH01] Hsiao-Wuen Hon Xuedong Huang, Alex Acero. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall, 2001.

[XZS16] Caiming Xiong, Victor Zhong e Richard Socher. Dynamic coattention networks for question answering. *CoRR*, abs/1611.01604, 2016.

Index

DFT, *see* transformada discreta de Fourier
DSP, *see* processamento digital de sinais

Fourier
    transformada, *see* transformada de Fourier

STFT, *see* transformada de Fourier de tempo reduzido

TBP, *see* periodicidade região codificante