Universidade Federal do Rio Grande do Norte
Centro de Ciências Exatas e da Terra
Departamento de Informática e Matemática Aplicada
Bacharelado em Ciência da Computação

Geração de prosódia para o português brasileiro em sistemas text-to-speech

Felipe Cortez de Sá

Natal-RN
Julho de 2018

Felipe Cortez de Sá

Geração de prosódia para o português brasileiro em sistemas text-to-speech

Monografia de Graduação apresentada ao Departamento de Informática e Matemática Aplicada do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de bacharel em Ciência da Computação.

Orientador

Dr. Carlos Augusto Prolo

Universidade Federal do Rio Grande do Norte – UFRN
Departamento de Informática e Matemática Aplicada – DIMAp

Natal-RN

Junho de 2018

Monografia de Graduação sob o título *Geração de prosódia para o português brasileiro em sistemas text-to-speech* apresentada por Felipe Cortez de Sá e aceita pelo Departamento de Informática e Matemática Aplicada do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Dr. Carlos Augusto Prolo
Orientador
Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte

Dr. Antônio Carlos Gay Thomé
Departamento de Informática e Matemática Aplicada
Universidade Federal do Rio Grande do Norte

Dra. Erica Reviglio Iliovitz
Departamento de Letras
Universidade Federal do Rio Grande do Norte

Natal-RN, 20 de junho de 2018.

Dedicado a várias pessoas

Agradecimentos

Obrigado várias pessoas

Some few people are born without any sense of time. As consequence, their sense of place becomes heightened to an excruciating degree. They lie in tall grass and are questioned by poets and painters from all over the world. These time-deaf are beseeched to describe the precise placement of trees in the spring, the shape of snow on the Alps, the angle of sun on a church, the position of rivers, the location of moss, the pattern of birds in a flock. Yet the time-deaf are unable to speak what they know. For speech needs a sequence of words, spoken in time.

Alan Lightman, Einstein's Dreams

Geração de prosódia para o português brasileiro em sistemas text-to-speech

Autor: Felipe Cortez de Sá
Orientador(a): Dr. Carlos Augusto Prolo

Resumo

Com a cada vez mais forte presença de smartphones e home assistants no cotidiano, grandes empresas de tecnologia vêm desenvolvendo sistemas de conversação baseados em fala, denominadas voice user interfaces. Apesar dos avanços, é perceptível que os sistemas de síntese de voz, especialmente para o português brasileiro, deixam a desejar quanto à naturalidade da fala gerada. Um dos fatores principais que contribuem para isso é a prosódia, isto é, entonação, ritmo e acento da fala. Este trabalho investiga sistemas text-to-speech existentes através do estudo de seus algoritmos para síntese de voz e geração de prosódia para diversas línguas, com foco no português brasileiro. São explicitados os desafios encontrados, é feito um levantamento de modelos de análise prosódica na linguística e propõem-se possíveis soluções para tornar a geração de voz mais próxima à humana.

Palavras-chave: text-to-speech, prosódia, voice user interfaces

Prosody generation for Brazilian Portuguese in text-to-speech systems

Author: Felipe Cortez de Sá
Advisor: Carlos Augusto Prolo, Ph.D.

Abstract

With the evergrowing presence of smartphones and home assistants in our daily lives, technology companies have been developing two-way conversation systems, that is, voice user interfaces. Despite its recent improvements, text-to-speech programs still sound artificial, especially for their Brazilian Portuguese voices. A big contributing factor for that is the lack of accurate prosody, that is, pitch, length and emphasis. This thesis explores existing text-to-speech systems, especially those for which there are Brazilian Portuguese voices, focusing on their prosody generation modules. We highlight challenges of prosody generation, review prosodic analysis in the Linguistics field and propose possible solutions for improving text-to-speech quality.

Keywords: text-to-speech, prosody, voice user interfaces

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

Lista de símbolos

Sumário

1 Introdução

Interfaces humano-computador que utilizam a voz, denominadas *voice user interfaces*, antigamente vistas apenas na ficção científica, hoje são uma realidade e estão disponíveis em *smartphones* e ambientes *desktop*. De acordo com (DUTOIT, 1997; JURAFSKY; MARTIN, 2009), há uma grande área de aplicação para essas interfaces, destacando-se a acessibilidade, permitindo que deficientes visuais possam ouvir texto sem a necessidade de gravação prévia de seu conteúdo. Além disso, com o aumento da popularidade de sistemas embarcados, é importante investigar novas formas de interação humano-máquina, e a síntese de fala, juntamente com o reconhecimento, permitem comunicação de duas vias com esses sistemas. Sistemas *text-to-speech* (doravante designados pela sigla TTS) também podem servir para pessoas que perderam a habilidade de falar, como o físico Stephen Hawking, que desde 1986 utilizou um sintetizador de voz para se comunicar, e o crítico de cinema Roger Ebert, que após perder a mandíbula passou a falar através de um sistema TTS, mais tarde usando uma solução personalizada que sintetizava uma aproximação de sua própria voz baseada em múltiplas gravações passadas.

Os serviços mais populares e robustos que temos atualmente são implementações proprietárias de grandes empresas, como Siri (Apple Inc., 2017), Cortana (Microsoft Corp., 2017) e Alexa (Amazon.com, Inc., 2017). Apesar da praticidade e ganho de acessibilidade providos por essas interfaces, os serviços disponíveis sintetizam voz com resultados perceptivelmente artificiais, principalmente para a língua portuguesa, se compararmos com os mesmos serviços para o inglês. Uma das causas da artificialidade é a prosódia empregada, isto é, o ritmo, entonação e acento. Mesmo com um sistema personalizado, Roger Ebert se queixava da falta de expressividade do algoritmo.

1.0.1 Objetivos

Neste trabalho propõe-se investigar as causas da artificialidade de prosódia em sistemas TTS disponíveis, estudando os modelos de prosódia, algoritmos e métodos para

síntese de fala para diversas línguas com foco no português, e implementar um ou mais modelos de geração de prosódia promissores para o português brasileiro baseados na estrutura sintática do texto de entrada identificada por técnicas de processamento de linguagem natural. Ao final da implementação, os resultados serão avaliados qualitativamente através de questionários, comparando-os ao estado da arte e disponibilizando o sistema publicamente.

1.1 Organização do trabalho

No capítulo 3 é feita uma revisão da literatura, revelando (?) os sistemas *text-to-speech* existentes tanto para o inglês quanto para o português brasileiro e como a prosódia é abordada em cada um deles. Mostramos como trabalhos recentes abordam síntese de fala. Também são descritos os trabalhos existentes em análise e síntese de prosódia em um contexto não necessariamente computacional.

No capítulo 4, justifica-se a abordagem escolhida para o sistema desenvolvido com base na revisão da literatura e descreve-se a implementação do software, incluindo sua arquitetura, as linguagens de programações utilizadas etc.

No capítulo 5, explicitamos os resultados, descrevendo a metodologia empregada na avaliação qualitativa das amostras de áudio geradas pelo *software* desenvolvido comparadas ao estado da arte. Ademais, propõem-se melhorias e trabalhos futuros que poderão ser realizados utilizando como base o que se desenvolveu nesta pesquisa.

2 Capítulo 2

2.1 TTS

2.1.1 Breve história?

Modelos físicos, Bell Labs.

2.1.2 Fonemas e fones

2.1.3 Abordagens

2.1.3.1 Difones

2.1.3.2 Unit selection

2.1.3.3 HMM

2.1.3.4 DNN

Resultados realistas, mas não há como controlar parâmetros

2.1.4 TTS em português

LianeTTS (MBROLA), HMM-based (Maia et al), MaryTTS (FalaBrasil)

2.2 Prosódia

2.2.1 Tipos de prosódia

2.2.1.1 Aumentativa

2.2.1.2 Suprasegmental

2.2.1.3 Afetiva

2.2.2 Elementos

Intonational tune Downdrift Microprosódia

2.2.3 Prosódia como elemento extra-textual

Justifica abordagem do trabalho: considerando o texto como sequência de palavras, é difícil determinar prosódia afetiva. Gerar a prosódia certa é uma questão de Natural Language Understanding, isto é, é preciso entender o texto para gerar os contornos melódicos afetivos.

2.2.4 Prosódia no português brasileiro

Trabalhos de Moraes, Tenani, ...

2.2.5 Modelos de prosódia

British school, autosegmental metrical, Fujisaki, Tilt, INTSINT. “The AM model is phonological, the INTSINT model phonetic and the Fujisaki and Tilt models acoustic”. INTSINT: IPA para prosódia (mais ou menos o que eu quero fazer, mas INTSINT é para análise). ref Moraes, Intonation Systems (20 languages). AM: Pierrehumbert, Moraes (pitch analysis by synthesis).

2.2.6 Trabalhos semelhantes?

3 Capítulo 3

(MIRANDA, 2015) analisa segundo o modelo IPO.

3.1 TTS

3.1.1 Breve história?

Modelos físicos, Bell Labs.

3.1.2 Fonemas e fones

3.1.3 Abordagens

3.1.3.1 Difones

3.1.3.2 Unit selection

3.1.3.3 HMM

3.1.3.4 DNN

Resultados realistas, mas não há como controlar parâmetros

3.1.4 TTS em português

LianeTTS (MBROLA), HMM-based (COUTO et al., 2010), MaryTTS (FalaBrasil) (Projeto Fala Brasil,).

3.2 Prosódia

3.2.1 Tipos de prosódia

3.2.1.1 Aumentativa

3.2.1.2 Suprasegmental

3.2.1.3 Afetiva

3.2.2 Elementos

Intonational tune Downdrift Microprosódia

3.2.3 Prosódia como elemento extra-textual

Justifica abordagem do trabalho: considerando o texto como sequência de palavras, é difícil determinar prosódia afetiva. Gerar a prosódia certa é uma questão de Natural Language Understanding, isto é, é preciso entender o texto para gerar os contornos melódicos afetivos.

3.2.4 Prosódia no português brasileiro

Trabalhos de Moraes, Tenani, ...

3.2.5 Modelos de prosódia

British school, autosegmental metrical, Fujisaki, Tilt, INTSINT. “The AM model is phonological, the INTSINT model phonetic and the Fujisaki and Tilt models acoustic”. INTSINT: IPA para prosódia (mais ou menos o que eu quero fazer, mas INTSINT é para análise). ref Moraes, Intonation Systems (20 languages). AM: Pierrehumbert, Moraes (pitch analysis by synthesis).

3.2.6 Trabalhos semelhantes?

4 Capítulo 4

4.1 Implementação

4.1.1 espeak-ng

Para realizar a normalização de texto, foi utilizado o programa *open-source* espeak-ng. Nele também é feita a conversão grafema-fonema (comumente representado pela sigla G2P). O resultado é passado para o programa desenvolvido neste trabalho.

4.1.2 MBROLA

Baseado no algoritmo PSOLA (Pitch Synchronous Overlap and Add) (descrever PSOLA) Fonemas/fones simplificados, determinados pelo autor de uma voz, gerando saída com voz sintetizada. Síntese por dífonos.

4.1.2.1 Formato

Pausas ou fones!!! e duração seguido por um ou mais pares de porcentagem e frequência (Hz)

4.1.3 Arquitetura

Diagrama aqui

4.1.4 Módulo de prosódia

4.1.4.1 Sintaxe

O programa foi codificado em Python em sua versão 3.6.

4.1.5 Sintaxe prosódica

A solução para melhorar a geração prosódica foi adicionar marcações à linguagem natural, denotando o contorno de acordo com o modelo modelo autossegmental e métrico Citar SSML (Speech Synthesis Markup Language).

4.1.6 Editor gráfico

Para alterar a prosódia manualmente com maior controle, foi desenvolvido um editor gráfico utilizando JavaScript. A duração e altura de cada fone pode ser especificado arrastando barras de controle. O editor se comunica com o espeak-ng e MBROLA através de um servidor programado em Python utilizando o *framework* Flask.

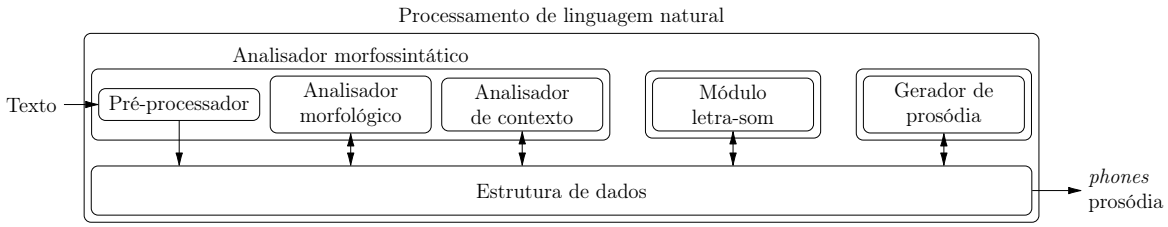

Figura 1: Diagrama

5 Capítulo 5

5.1 Resultados

5.1.1 Metodologia

5.1.2 MOS

5.2 Trabalhos futuros

6 Considerações finais

As considerações finais formam a parte final (fechamento) do texto, sendo dito de forma resumida (1) o que foi desenvolvido no presente trabalho e quais os resultados do mesmo, (2) o que se pôde concluir após o desenvolvimento bem como as principais contribuições do trabalho, e (3) perspectivas para o desenvolvimento de trabalhos futuros. O texto referente às considerações finais do autor deve salientar a extensão e os resultados da contribuição do trabalho e os argumentos utilizados estar baseados em dados comprovados e fundamentados nos resultados e na discussão do texto, contendo deduções lógicas correspondentes aos objetivos do trabalho, propostos inicialmente.

Referências

Amazon.com, Inc. *Alexa*. 2017. Disponível em: <https://developer.amazon.com/alexa>.

Apple Inc. *Siri*. 2017. Disponível em: <https://www.apple.com/ios/siri/>.

COUTO, I. et al. An open source hmm-based text-to-speech system for brazilian portuguese. In: *7th international telecommunications symposium*. [S.l.: s.n.], 2010.

DUTOIT, T. *An introduction to text-to-speech synthesis*. [S.l.: s.n.], 1997. (Text, Speech and Language Technology 3).

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009. ISBN 0131873210. Disponível em: <https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf>.

Microsoft Corp. *Cortana*. 2017. Disponível em: <https://www.microsoft.com/en-us/windows/cortana>.

MIRANDA, L. *Análise da entoação do português do Brasil segundo o modelo IPO*. Tese (Doutorado) — Dissertação de mestrado em Língua Portuguesa. Rio de Janeiro: UFRJ, 2015.

Projeto Fala Brasil. Disponível em: <http://www.laps.ufpa.br/falabrasil/descricao.php>.

APÊNDICE A – Primeiro apêndice

Os apêndices são textos ou documentos elaborados pelo autor, a fim de complementar sua argumentação, sem prejuízo da unidade nuclear do trabalho.

ANEXO A – Primeiro anexo

Os anexos são textos ou documentos não elaborado pelo autor, que servem de fundamentação, comprovação e ilustração.