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
Junho 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

TTS – *text-to-speech*

INTSINT – *International Transcription System for Intonation*

HMM – *Hidden Markov Model*

SSML – *Speech Synthesis Markup Language*

XML – *eXtensible Markup Language*

W3C – *World Wide Web Consortium*

MBRPSOLA – *Multi-Band Resynthesis Pitch Synchronous Overlap and Add*

Lista de símbolos

% (fronteira de enunciado para ToBI)

ms (milissegundos)

Hz (Hertz)

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 TTS) também podem auxiliar 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., 2011), Cortana (Microsoft Corp., 2014) e Alexa (Amazon.com, Inc., 2014). 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. 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.

1.1 Objetivos

Este tem o objetivo de fazer um levantamento do estado da arte de algoritmos para TTS, com atenção especial a como a prosódia é abordada em cada um deles, e propor melhorias possíveis para módulos de prosódia para o português brasileiro.

1.2 Organização do trabalho

No capítulo 2, é feita uma breve explanação dos conceitos abordados neste trabalho, com foco em TTS e prosódia.

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

No capítulo 4, é descrito o sistema desenvolvido, justificando a abordagem com base na revisão da literatura e explica-se a implementação do software, incluindo sua arquitetura, as linguagens de programações utilizadas e o funcionamento.

No capítulo 5, apresentamos os resultados obtidos com a implementação.

2 Fundamentação teórica

2.1 Prosódia

2.1.1 Componentes

Taylor (2009) fala que a prosódia tem: Stress (loudness and phonatory force) Syllabic length F0, intensidade, duração Intonational tune Downdrift Microprosódia

2.1.2 Função prosódica

Taylor (2009) mostra três razões pelas quais utilizamos prosódia na fala.

2.1.2.1 Afetiva

Prosódia pura, emoção, estado mental e atitude.

2.1.2.2 Suprassegmental

Características em um discurso neutro, isto é, quando uma sentença é falada sem conteúdo afetivo significante. Taylor (2009) fala que em seu modelo de prosódia a parte supra-segmental não é conteúdo prosódico verdadeiro, mas sim um outro aspecto do componente verbal.

2.1.2.3 Aumentativa

Um elemento para assegurar a comunicação efetiva do componente verbal de uma mensagem.

2.1.3 Prosódia como elemento extra-textual

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 Sistemas TTS

Um sistema *text-to-speech* converte texto em fala.

2.2.1 Estrutura

Na literatura, encontramos diversas arquiteturas específicas de um sistema TTS. Dutoit (1997) propõe um diagrama geral, dividindo sistemas em duas partes principais:

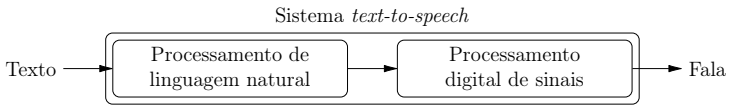

Figura 1: Arquitetura geral de sistemas TTS (adaptado de Dutoit (1997))

É comum encontrar em outros trabalhos o termo *front end* para o bloco de processamento de linguagem natural e *back end* para o bloco de processamento digital de sinais. Doravante utilizaremos essa nomenclatura.

2.2.2 *Front end*

2.2.2.1 Normalização de texto

O texto de entrada pode conter números e símbolos. A primeira parte do processo de conversão de texto para fala é transformar em sua representação por extenso. Sentence tokenization, isto é, descobrir as boundaries de cada sintagma. Depois, normalização de palavras não-padrão? abreviações, acrônimos e siglas. Desambiguação de homônimos heterófonos (e.g.: "Gosto de pão"). Análise fonética (normalized word strings para pronúncia).

2.2.2.2 Conversão grafema-fone

Palavras não-padrão são colocadas num dicionário de pronúncia. O resto é calculado de acordo com regras letra-som. Context-dependent e independent. Abordagens por machine

learning ou regras. A conversão grafema-fone (ou fonema) consiste em transformar o texto normalizado em fones, ou seja, uma sequência de caracteres em uma sequência de fones.

2.2.2.3 Geração de prosódia

A partir do texto e fones gerados nas etapas anteriores, deve-se estimar uma curva F0, ritmo e ênfase.

2.2.3 Back end

Com o texto de entrada transformado em fones e informação prosódica, um *back end* é responsável por gerar uma forma de onda. Jurafsky e Martin (2009) separam algoritmos de síntese em três paradigmas: síntese concatenativa, síntese por formantes e síntese articulatória.

2.2.3.1 Síntese articulatória

2.2.3.2 Síntese por formantes

2.2.3.3 Síntese concatenativa

Síntese concatenativa pode trabalhar com diferentes atomicidades: nível de palavra, difonos e fones individuais.

3 Revisão da literatura

3.0.1 Abordagens

3.0.1.1 Klaat

Síntese por formantes. (DUNN, 2006) usa uma mistura do algoritmo de Klatt com sons de consoantes pré-gravados.

3.0.1.2 *Unit selection* e dífonos

Abordagem utilizada pelo programa MBROLA. Consiste em gravar fala, separar pedaços de dois em dois.

3.0.1.3 HTS

Algumas vozes para o MaryTTS (SCHRöDER; TROUVAIN, 2003) utilizam HMMs, isto é, Modelos ocultos de Markov para *unit selection*. Há, inclusive, uma voz brasileira feita a partir de HMMs: (COUTO et al., 2010). Projetos mais recentes como (WU; WATTS; KING, 2016; ZE; SENIOR; SCHUSTER, 2013) utilizam redes neurais para estimação de parâmetros. O trabalho da Google informa? que a estimação da curva F0 é uma possível melhoria.

3.1 Sistemas TTS

3.1.1 MaryTTS

(COUTO et al., 2010), MaryTTS.

3.1.2 LianeTTS

Projeto da SERPRO LianeTTS. Utiliza MBROLA como *back end*. A geração de prosódia é simples, baseada em tabela.

3.1.3 espeak

Projeto do Dunn. Síntese por formantes (com consoantes gravadas). É possível obter apenas até a parte de g2p como saída.

3.1.4 Festival

Edinburgo. Multilingual, mas não pro português. Modular: permite configuração independente de subsistemas como regras letra-som, POS tagging, entonação. Suporte a múltiplos sintetizadores de fala. HTS, Unit selection, MBROLA.

3.1.5 Microsoft

Braga et al. (2008) descrevem sistema TTS desenvolvido pela Microsoft baseado em HMM. A stochastic-based LTS (Letter-to-Sound) converter to predict phonetic transcriptions for out-of-vocabulary words and the prosody models, which are data- driven using a prosody tagged corpus of 2000 sentences.

3.1.6 Aiuruetê

Usa PSOLA pra recodar unidades armazenadas (ortofones)

3.1.7 LINSE-DNN

statistical parametric speech synthesis (MAIA; SEARA, 2017)

3.1.7.1 MBROLA

Baseado em síntese por dífonos. Antigo mas qualidade decente. Três vozes para o português brasileiro. (DUTOIT et al., 1996)

Baseado no algoritmo MBRPSOLA (DUTOIT; LEICH, 1993). É um *back-end*. O programa recebe uma lista de fones. Um exemplo de entrada é:

```
_ 150 50 150
o 108 50 125
l 125 50 75
a 116 20 232 80 300
_ 150 50 150
```

3.1.7.2 Formato

Em cada linha, tem-se um fone ou um silêncio representado pelo *underscore* seguido por uma duração em milissegundos e, por último, um ou mais pares de porcentagem e frequência em Hertz determinando alvos para a curva F0. Como exemplo, na quarta linha temos o fone a com duração de 116 ms e dois alvos para altura, 232 Hz em 20% e 300 Hz em 80%.

Cada voz gravada provê uma tabela com os fones que podem ser utilizados. Utilizamos neste trabalho a voz br3 desenvolvida por Denis R. Costa disponível no site oficial do projeto MBROLA.

3.1.8 Prosódia em sistemas TTS

3.1.8.1 SSML

SSML foi criado para padronizar anotações em sintetizadores de fala, permitindo desambiguação de pronúncia, mudança de linguagem, etc. Phrasing, emphasis, pronunciation. É baseado em XML. (BURNETT; SHUANG, 2010), A especificação é mantida pela W3C. SSML é apenas um modificador (insuficiente para estimar prosódia?) Anotações em SSML são úteis para estender o texto original e dar dicas para sintetizadores de fala. A especificação cita os elementos *emphasis*, *break* e *prosody* como marcadores que podem auxiliar o processador de linguagem natural a gerar parâmetros prosódicos apropriados. MaryTTS usa ToBI, MBROLA, Unit selection do FreeTTS, SSML. Suportado (?) pelo MaryTTS e sistemas proprietários da Alexa, Google Assistant e Cortana.

```
<speak>
    Eu quero
    <emphasis level="strong">aquele</emphasis>
    carro.
</speak>
```

3.1.8.2 EmotionML

EmotionML (SCHRÖDER; BURKHARDT, 2014) foi criada para várias coisas, uma delas é ajudar algoritmos a determinarem prosódia. (CHARFUELAN; STEINER, 2013) descreve um *framework* para implementação de determinação de prosódia a partir de anotações em emotionml.

3.1.9 Modelos de prosódia

3.1.9.1 Teoria métrica-autossegmental

Pierrehumbert, Moraes (pitch analysis by synthesis). ref Moraes, Intonation Systems (20 languages).

3.1.9.2 DaTo

3.1.9.3 IPO

(MIRANDA, 2015) analisa a prosódia para o português brasileiro através do modelo IPO, seguindo crítica de Lucente que o sistema ToBI é inapropriado.

3.1.9.4 INTSINT

INTSINT é um sistema de anotação para prosódia.

4 Editor de prosódia

4.1 Implementação

4.1.1 espeak-ng

Foi utilizado o programa *open-source* espeak-ng (DUNN, 2006) para realizar a normalização de texto e realizar a conversão grafema-fonema, ou seja, obter a partir do texto de entrada uma representação em fonemas. O resultado é passado para o programa desenvolvido neste trabalho.

Apesar da existência de outras ferramentas para *front-end* para o português brasileiro, optamos por esta pela facilidade de instalação, marcação de ênfase disponível, etc.

4.1.2 Arquitetura

Diagrama aqui

4.1.3 Módulo de prosódia

4.1.4 Editor gráfico

O programa foi codificado em Python em sua versão 3.6. Pega resultado do espeak-ng, processa com editor gráfico e gera MBROLA.

Para alterar a prosódia manualmente, foi desenvolvido um editor gráfico para *web* utilizando HTML, CSS e 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 para prover *endpoints* de uma API REST.

Figura 2: Arquitetura do sistema desenvolvido

5 Resultados

Necessidade de um corpus anotado com prosódia para o português.

6 Considerações finais

Neste trabalho apresentamos a definição de TTS, fizemos uma busca dos sistemas open-source existentes investigamos como a prosódia funciona nos sistemas encontrados especialmente para o português brasileiro investigamos como melhorar passos para o futuro implementação básica de uma provável melhoria

6.1 Trabalhos futuros

Usar técnicas de Natural Language Understanding para gerar notação EmotionML automaticamente. Desenvolvimento de corpus anotados com prosódia para o português brasileiro.

Referências

Amazon.com, Inc. *Alexa*. 2014. Disponível em: <https://developer.amazon.com/alexa>. Acesso em 25 de setembro de 2017.

Apple Inc. *Siri*. 2011. Disponível em: <https://www.apple.com/ios/siri/>. Acesso em 25 de setembro de 2017.

BRAGA, D. et al. HMM-based Brazilian Portuguese TTS. *Braga et al.(eds), Propor*, p. 47–50, 2008.

BURNETT, D.; SHUANG, Z. W. *Speech Synthesis Markup Language (SSML) Version 1.1*. [S.l.], 2010. Disponível em: <http://www.w3.org/TR/2010/REC-speech-synthesis11-20100907/>. Acesso em 5 de junho de 2018.

CHARFUELAN, M.; STEINER, I. Expressive speech synthesis in MARY TTS using audiobook data and emotionML. In: *INTERSPEECH*. [S.l.: s.n.], 2013. p. 1564–1568.

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.

DUNN, R. H. *espeak-ng*. 2006. Disponível em: <https://github.com/espeak-ng/espeak-ng>. Acesso em 29 de outubro de 2017.

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

DUTOIT, T.; LEICH, H. MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, Elsevier, v. 13, n. 3-4, p. 435–440, 1993.

DUTOIT, T. et al. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In: IEEE. *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. [S.l.], 1996. v. 3, p. 1393–1396.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009. ISBN 0131873210.

MAIA, R.; SEARA, R. Um sistema TTS baseado em redes neurais profundas usando parâmetros síncronos de pitch. p. 388–392, 2017.

Microsoft Corp. *Cortana*. 2014. Disponível em: <https://www.microsoft.com/en-us/windows/cortana>. Acesso em 25 de setembro de 2017.

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.

SCHRÖDER, M.; BURKHARDT, F. *Emotion Markup Language (EmotionML) 1.0*. [S.l.], maio 2014. Disponível em: <http://www.w3.org/TR/2014/REC-emotionml-20140522/>. Acesso em 8 de junho de 2018.

SCHRöDER, M.; TROUVAIN, J. The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, Springer, v. 6, n. 4, p. 365–377, 2003.

TAYLOR, P. *Text-to-speech synthesis*. [S.l.]: Cambridge University Press, 2009.

WU, Z.; WATTS, O.; KING, S. Merlin: An open source neural network speech synthesis system. *Proc. SSW, Sunnyvale, USA*, 2016.

ZE, H.; SENIOR, A.; SCHUSTER, M. Statistical parametric speech synthesis using deep neural networks. In: IEEE. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. [S.l.], 2013. p. 7962–7966.