



Universidade de Brasília

**Instituto de Ciências Exatas
Departamento de Ciência da Computação**

Síntese de Voz

Leandro Ramalho Motta Ferreira

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Orientador
Prof. Dr. Jorge Carlos Lucero

Brasília
2016

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Curso de Computação — Licenciatura

Coordenador: Prof. Dr. Coordenador

Banca examinadora composta por:

Prof. Dr. Jorge Carlos Lucero (Orientador) — CIC/UnB
Prof. Dr. Professor I — CIC/UnB
Prof. Dr. Professor II — CIC/UnB

CIP — Catalogação Internacional na Publicação

Ferreira, Leandro Ramalho Motta.

Síntese de Voz / Leandro Ramalho Motta Ferreira. Brasília : UnB, 2016.
35 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2016.

1. Síntese, 2. Voz, 3. Saúde

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília-DF — Brasil



Universidade de Brasília

**Instituto de Ciências Exatas
Departamento de Ciência da Computação**

Sintese de Voz

Leandro Ramalho Motta Ferreira

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Prof. Dr. Jorge Carlos Lucero (Orientador)
CIC/UnB

Prof. Dr. Professor I Prof. Dr. Professor II
CIC/UnB CIC/UnB

Prof. Dr. Coordenador
Coordenador do Curso de Computação — Licenciatura

Brasília, 10 de maio de 2016

Dedicatória

Dedico a....

Agradecimentos

Agradeço a....

Resumo

AINDA Não tem

Palavras-chave: Sintese, Voz, Saúde

Abstract

Still there isn't.

Keywords: Synthesis, Voice, Health

Sumário

1	Introdução	1
1.1	Motivação	1
2	Conceitos Básicos de Síntese de Voz	2
2.1	Anatomia da Voz	2
2.1.1	Órgãos	2
2.1.2	Aparelho Fonador	2
2.2	Propriedades Físicas	2
2.2.1	Lei Bernoulli	2
2.2.2	Crítérios para Oscilação	3
2.2.3	Tipos de Oscilação	3
2.2.4	Tensão	3
2.2.5	Curva Força e Alongamento	3
2.2.6	Viscosidade	4
2.2.7	Reflexão de Som	4
2.3	Sintetizadores de Voz	4
2.3.1	Modelo Massa-Mola Auto Sustentável	4
2.3.2	Synpath	5
2.3.3	HMMs	5
2.3.4	MHRMM	5
2.3.5	FrameWorks Sintetizador de Voz	6
2.3.6	Envoltória F0	6
2.3.7	Síntese de Voz em Mandarim	7
2.4	Sistema Auditivo	7
2.4.1	Introdução	7
2.4.2	Intensidade	7
2.5	Voz e Propriedades Linguísticas	7
2.5.1	Vogais	8
2.5.2	Consoantes	8
	Referências	9

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Capítulo 2

Conceitos Básicos de Síntese de Voz

2.1 Anatomia da Voz

2.1.1 Órgãos

Órgãos envolvidos com a síntese da voz:

- Diafragma
- Pulmões
- Traqueos
- Esôfago
- laringe
- epiglote
- faringe
- palato , língua cavidade nasal, dente lábios(7)

2.1.2 Aparelho Fonador

O estudo do aparelho fonador começa-se por suas estruturas e componentes importantes. Após um estudo detalhado dos fenômenos físicos e como se comportam é essencial também.

2.2 Propriedades Físicas

2.2.1 Lei Bernoulli

Energia potencial e energética cinética em fluidos se mantém a mesma porém em proporções diferentes (13):

$$P + \frac{\rho * v^2}{2} = \text{Constante}$$

Sendo : ρ = Densidade do fluido
 P = Pressão no duto onde o fluido se encontra
 v = velocidade

2.2.2 Critérios para Oscilação

Alguns critérios devem ser atendidos para que um determinado padrão de movimento seja considerado como uma oscilação mecânica, a saber:

No sistema onde ocorre o movimento deve haver uma posição de equilíbrio estável, que é caracterizada por uma força restaurativa que sempre acelera o corpo em movimento de volta para a sua posição de repouso. Deve haver inércia (no caso do sistema mecânico, a massa atua como propriedade de inércia) no sistema para superar esta posição de equilíbrio. A perda, em excesso, de energia por ciclo de oscilação deve ser zero...

2.2.3 Tipos de Oscilação

De acordo com TITZE (14), os tipos de oscilação são:

- Oscilação Natural: Quando um sistema que se encaixa nos critérios anteriores se move sem interferência após um distúrbio inicial.
- Oscilação Natural: Quando um sistema que se encaixa nos critérios anteriores se move sem interferência após um distúrbio inicial.
- Oscilação Forçada: Requer uma fonte externa de condução que por si só é um oscilador. Dita grande parte do padrão de vibração do sistema.
- Oscilação Auto-Sustentável: Requer uma fonte de energia estável e uma interação não-linear entre os componentes internos ao sistema. As perdas de energia são compensadas, mantendo o padrão oscilatório.

2.2.4 Tensão

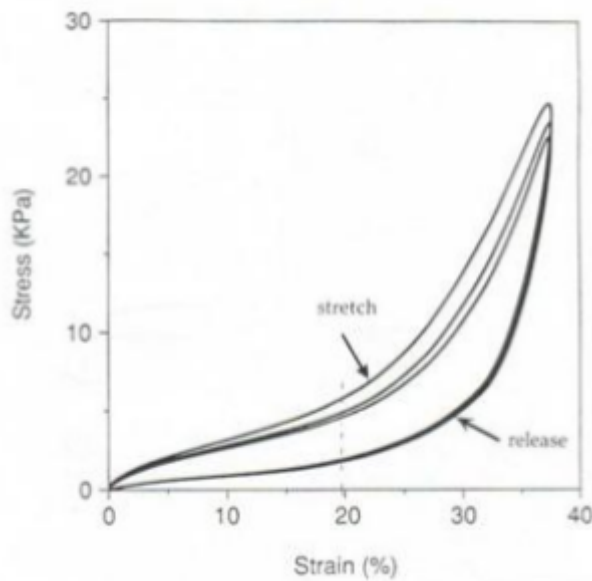
Conceito de Tensão For por unidade de ar

$$\sigma = \frac{F}{A}$$

Sendo: f : força aplicada.
A: área de aplicação desta força.

2.2.5 Curva Força e Alongamento

Utiliza-se para não ser dependente da geometria do material. Utilizamos nas cordas vocais(?) por serem materiais biológicos. Cria uma figura ilustrando comportamento da deformação das pregas vocais



2.2.6 Viscosidade

É a velocidade de deformação (consequentemente, de restauração) de um determinado fluido quando atuam forças de tensão no mesmo. Matematicamente pode ser expresso conforme a equação seguinte: ?

$$\sigma = \eta * \frac{d\epsilon}{dt}$$

2.2.7 Reflexão de Som

Um fenômeno ligado a rigidez e amortecimento entre um meio e outro. (7) Ondas quando tentam penetrar em um segundo meio, sendo o segundo meio rígido, as partículas do primeiro meio se aglomeram tentando passar porém falham, seu acúmulo de partículas gera pressão que acaba criando uma outra onda no primeiro meio decorrente da primeira onda. (4)

O mesmo ocorre com o meio 2 sendo totalmente não rígido e o primeiro meio sendo bem rígido, exaurindo excesso de partículas do meio 1 no meio 2 criando rarefação no meio 1, o que cria uma outra onda de pressão negativa (1). A propagação é sempre em direção oposta à fonte, no caso é na direção contrária à coluna de ar (meio 1).

2.3 Sintetizadores de Voz

2.3.1 Modelo Massa-Mola Auto Sustentável

: A fechadura e abertura da glote num sistema massa mola de apenas um lado

$$P = \left(1 - \frac{a_2}{a_1}\right) * (P_s - P_i) + P_i$$

Versão simplificada da pressão massa mola

- P: Pressão resultante NA GLOTE
- a1: Areas de entrada da glote
- a2: Areas de saída da glote
- Ps: Pressão subglótica
- Pi: Pressão sobre o trato vocal(Pressão input)

No modelo mono massa $a1 = a2$. No caso em que Pressão na GLOTE, P, seja igual a pressão supraglotal, indica que a inércia da coluna de ar acima da glote altera a pressão.

2.3.2 Synpath

Sintetizador computacional, criado por Lucero. Sintetizador concebido por Fraj (3) Foi escolhido não usar o modelo multi-massa por variações não suaves e é tematicamente muito complexo apesar instabilidades numéricas. Synpath utiliza como base de representação das pregas vocais o modelo mono massa. O trato vocal é representado por uma concatenação de tubos cilíndricos os quais se propagam uma onda acústica. Superglótica não leva em conta o trato vocal. Perdas e viscosidade são considerada.

2.3.3 HMMs

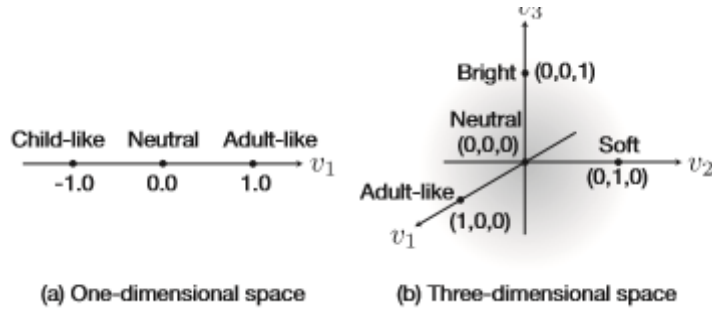
Minera-se de várias partituras musicais para treino. Os dados minerados dessas músicas são fonemas, altura, intensidade e os intervalos, isto é relação com outras notas. Esses dados são convertidos e mapeados em "labels" dependentes de contexto (2). Após isso as HMM's são treinados através dos dados de treinamento usando o algoritmo EM.(5). Após isso ocorre a fase de síntese, usa-se outra partitura para ser convertida em "labels" dependentes de contexto e estima-se quais "labels" pré-processadas são correspondentes.(10)

2.3.4 MHRSMM

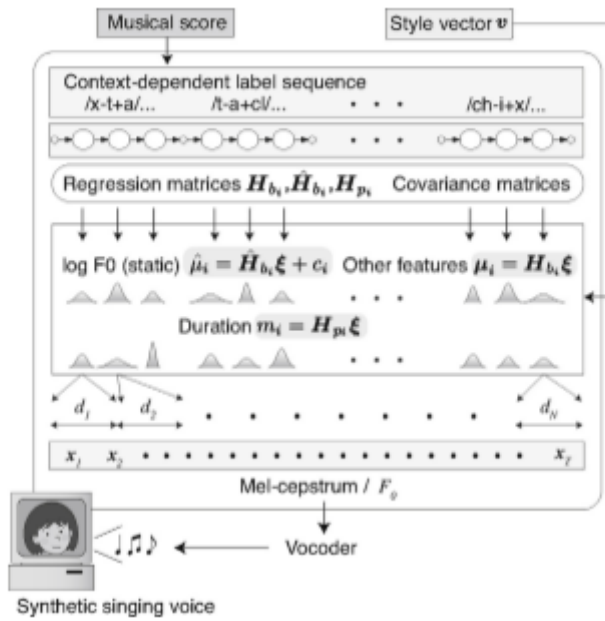
Uma variação HSMM. Modelo de múltipla regressão HSMM. Parâmetros importantes são μ_i e m_i dos outputs pdfs

$$\begin{aligned}\mu_i &= H_{bi}\xi \\ m_i &= H_{pi}\xi \\ \xi &= [1, v_1, v_2, \dots, v_L]^T \\ \xi &= [1, v^T]^T\end{aligned}$$

Onde L é a dimensão do vetor de estilo e v_i é a intensidade do i -ésimo estilo de canto. Um exemplo de um vetor de estilos de canto de tamanho $L = 2$ e $L = 3$.



Controle do Sintetizador de voz cantada baseado em MRHSMM



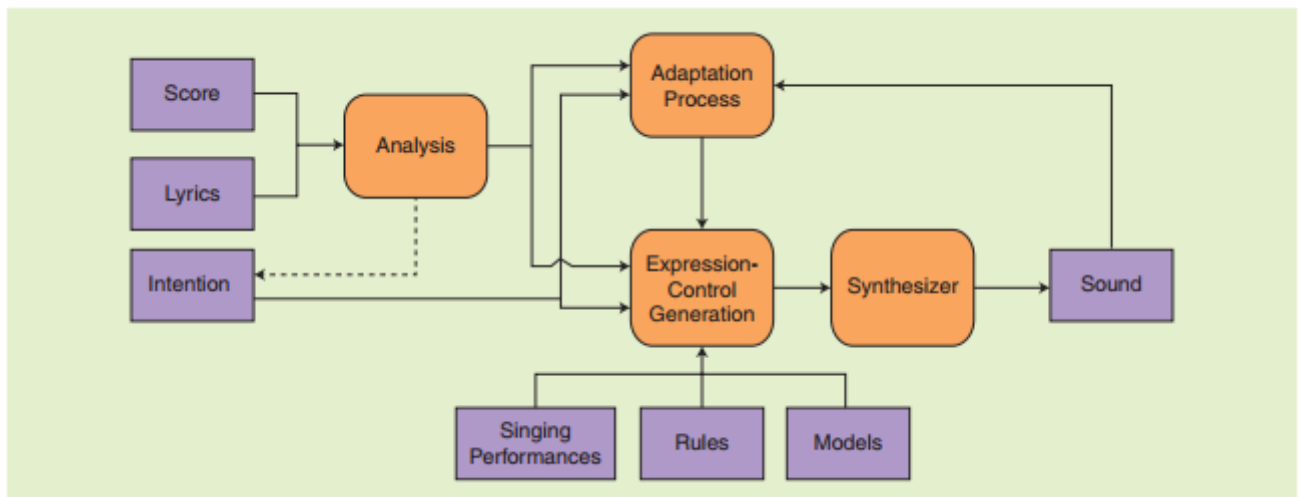
Durante a fase de síntese o usuário do programa adiciona vetores de estilos de acorodo com a intenção e a expressividade pretendida. Parametros de output como duração são gerados pelos vetores de estilos dados e matrizes de regressões treinadas usando MRHSMMs

Resultado de todo esse processo é um sequência HSMM usando parametros de geração de fala

O processo todo pode ser resumido abaixo

2.3.5 FrameWorks Sintetisador de Voz

Frame Work de um Sistema Sintetisador de Voz:



Input

Consiste da partitura, letra e emoção. o input é analisado e derivado em uma transcrição fonética, alinhamento com a performance alvo ou dados contextuais.(15)

Expressão

Expressão musical é um conceito intuitivo porém difícil de se definir. A expressão é chave na percepção da qualidade e naturalidade musical. No caso da voz cantada implica-se usar vários outros parâmetros além de frequência e amplitude. Psicologicamente contornos do timbre, vibrato, tremolo, timing fonético.(15)

2.3.6 Envoltória F0

Envoltórias F0 são usadas para expressar informação linguística, para-linguística e não-linguística.(11)

As Envoltórias F0 apresentam três (3) características importantes que fazem diferenciar uma voz falada de uma voz cantada.(12)

- 1 - O alcance dinâmico de uma envoltória F0 é mais largo que o de uma voz falada
- 2 - A envoltória F0 corresponde e tende a se manter estável em uma nota. A mudança de nota de uma envoltória F0 corresponde a melodia da música
- 3 - Existem muitas flutuações f0 que são apenas observadas em apenas vozes cantadas

2.3.7 Síntese de Voz em Mandarim

Utiliza-se a técnica HNM para a síntese da voz cantada em mandarim. HNM significa, "harmonic plus noise model". O modelo HNM divide o espectro de um sinal em dois(2) com larguras não iguais para modelagem melhor do espectro.(9)

2.4 Sistema Auditivo

2.4.1 Introdução

O Sistema auditivo consiste em componentes periféricos e centrais. Atualmente a maior parte do conhecimento do funcionamento dos sistemas auditivos deriva de estudos de animais não humanos.(8)

Sistema auditivo diferencia-se entre espécies em jeitos interessantes. Por exemplo algumas espécies tem características diferentes relacionadas aos sinais vocais mais utilizados por ela mesma.(8)

2.4.2 Intensidade

Como a frequência de um estímulo a intensidade dele é processado e codificado sub-cortical nos dois lados do cérebro em todos os níveis no cérebro.(8)

2.5 Voz e Propriedades Linguísticas

Uma divisão importante de acordo com Flanagan, são as letras separados em classificações Vogais e Consoantes que se associam a um movimento do trato vocal correspondente (6).

Place of articulation	Voiced		Voiceless	
Labio-dental	/v/	vote	/f/	for
Dental	/ð/	then	/θ/	thin
Alveolar	/z/	zoo	/s/	see
Palatal	/ʒ/	azure	/ʃ/	she
Glottal			/h/	he

2.5.1 Vogais

O trato vocal ao produzir uma vocal, em um articulação normal, mantém-se relativamente estável. Há uma opção de contribuição das cavidades nasais, uma cobertura porém é negligenciável. Baseado nessas características é divididos todas as consoantes. A tabela abaixo explica:

2.5.2 Consoantes

Sons produzidos com constrictões em algum ponto no trato vocal. Dividido em quatro (4) classes, baseados em duas funcionalidades binárias, sonorant e continuant.

Sonorant

Sonorant pode ser traduzido como "cantado". Consoantes Sonorant são sons que não aumentam a pressão do ar dentro do trato vocal pois a constrição não é muito justa ou o palato continua aberto, deixando ar escapar por ele.

Continuant

Uma consoante discontinuant é produzida por um fechamento completo em algum ponto no trato vocal.

Referências

- [1] Ferdinand de Saussure. *Self-oscillating source for vocal-tract synthesis*. IEEE Tran. Audio Eletroacoust, Audio Eletroacoust., 1968. 4
- [2] Gilles Degottex, Pierre Lanchantin, Axel Roebel, and Xavier Rodet. Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55(2):278 – 294, 2013. 5
- [3] Jean Schoentgen Samia Fraj and Francis Grenez. *Development and perceptual assesment of a synthesizer of disordered voices*. Journal of Acoustical Society of America, 2012. 5
- [4] Henry Gray. *Anatomy of the Human Body*. IEEE Tran. Audio Eletroacoust. 4
- [5] Jon Gudnason, Mark R.P. Thomas, Daniel P.W. Ellis, and Patrick A. Naylor. Data-driven voice source waveform analysis and synthesis. *Speech Communication*, 54(2):199 – 211, 2012. 5
- [6] James L. Flanagan Jont B. Allen Mark A. Hasegawa-Johnson. *Speech Analysis Synthesis and Perception*. 2008. 7
- [7] Matias Zanartu Kelley C. Stewart Michael W. Plesniak David E. Sommer Sean D. Peterson Byron D. Erath. *A review of lumped-element models of voiced speech*. Speech Communication. 2, 4
- [8] Jody Kreiman and Diana Van Lancker Sidtis. *Foundations of voice studies*, 2011. 7
- [9] Chyi-Yeu Lin, Li-Chieh Cheng, Chang-Kuo Tseng, Hung-Yan Gu, Kuo-Liang Chung, Chin-Shyurng Fahn, Kai-Jay Lu, and Chih-Cheng Chang. A face robot for autonomous simplified musical notation reading and singing. *Robotics and Autonomous Systems*, 59(11):943 – 953, 2011. 7
- [10] Takashi Nose, Misa Kanemoto, Tomoki Koriyama, and Takao Kobayashi. Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling. *Computer Speech Language*, 34(1):308 – 322, 2015. 5
- [11] Takeshi Saitou, Masashi Unoki, and Masato Akagi. Development of an {F0} control model based on {F0} dynamic characteristics for singing-voice synthesis. *Speech Communication*, 46(3–4):405 – 417, 2005. Quantitative Prosody Modelling for Natural Speech Description and Generation International Conference on Speech Prosody. 6

- [12] Takeshi Saitou, Masashi Unoki, and Masato Akagi. Development of an $\{F_0\}$ control model based on $\{F_0\}$ dynamic characteristics for singing-voice synthesis. *Speech Communication*, 46(3–4):405 – 417, 2005. Quantitative Prosody Modelling for Natural Speech Description and Generation International Conference on Speech Prosody. 7
- [13] BradH Story. *Tubetalker*. Dept. of Speech, Language, and Hearing, Tucson, AZ, 2010. 2
- [14] Ingo Titze. *Principles of Voice Production*. Prentice Hall, New Jersey 07632, 1994. 3
- [15] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg. Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. *IEEE Signal Processing Magazine*, 32(6):55–73, Nov 2015. 6