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

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.

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

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

1. Sintese, 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
CIC/UnB

Prof. Dr. Professor II
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 Sintese de Voz **2**
2.1 Anatomia da Voz 2
2.1.1 Aparelho Fonador 2
2.1.2 Músculos e Cartilagens 3
2.2 Propriedades Físicas 4
2.2.1 Lei Bernouli 4
2.2.2 Critérios para Oscilação 4
2.2.3 Tipos de Oscilação 4
2.2.4 Tensao 4
2.2.5 Curva Força e Alongamento 5
2.2.6 Viscosidade 5
2.2.7 Reflexao de Som 5
2.2.8 Fluxo de Ar na Glote 5
2.3 Sintetizadores de Voz 6
2.3.1 Modelo Massa-Mola Auto Sustentável 6
2.3.2 Synpath 6
2.3.3 HMMs 7
2.3.4 MHRSMM 8
2.3.5 FrameWorks Sintetisador de Voz 9
2.3.6 Envoltoria F0 9
2.3.7 Sintese de Voz em Mandarím 10
2.4 Sistema Auditivo 10
2.4.1 Introdução 10
2.4.2 Intensidade 10
2.5 Voz e Propriedades Linguísticas 10
2.5.1 Vogais 11
2.5.2 Consoantes 11

Referências **12**

Lista de Figuras

2.1 Aparelho Fonador 3
2.2 Secção coronal da laringe e parte superior da traquéia 3

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Capítulo 2

Conceitos Básicos de Sintese de Voz

2.1 Anatomia da Voz

Para estudar a produção e a síntese da voz, é necessário ter um conhecimento acerca da anatomia e do funcionamento físico da voz 10. Sendo assim, as subseções seguintes descreverão brevemente detalhes da anatomia do sistema fonador humano e como o som é produzido, moldado e inuenciado por este sistema.

2.1.1 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.

A Figura 2.1 10, mostra os órgãos associados com a produção da voz.

Dentro das condições normais, a voz é produzida quando um uxo de ar vindo dos pulmões é convertido em energia acústica através da vibração das pregas vocais, localizadas na laringe. Os padrões de vibrações resultantes são moldados acusticamente quando o som passa pelo trato vocal acima da laringe. O sistema respiratório serve como uma fonte de potência para a produção do som, sendo responsável por movimentar o ar através do trato vocal. A laringe atua como um oscilador convertendo a potência aerodinâmica produzida em energia sonora, sendo frequentemente retratada como a fonte da voz. No entanto, a mais importante função da laringe não é a produção de som, e sim, vedar as vias aéreas aos pulmões completamente, protegendo-as de objetos estranhos ou líquidos, principalmente durante a deglutinação. De maneira análoga, a laringe serve como uma válvula de acesso às vias respiratórias e por essa característica, atua também no controle do uxo de ar que por elas passam. Sendo assim,é fácil notar que há uma necessidade de mobilidade para toda estruturada laringe, logo é de se esperar que sua estrutura seja formada em sua maioria por cartilagens. De fato o é, com exceção de um osso chamado de Hioide, a laringe é basicamente formada por cartilagens e músculos. A seguir, analisaremos brevemente a dinâmica dos músculos e cartilagens da laringe.

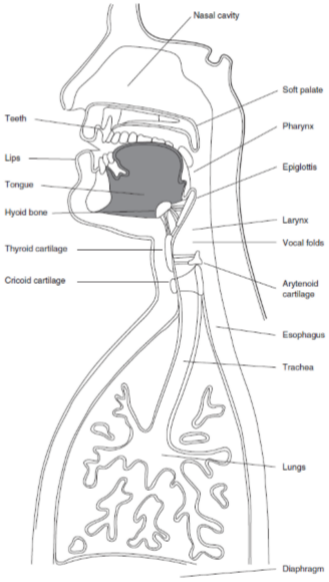


Figura 2.1: Aparelho Fonador

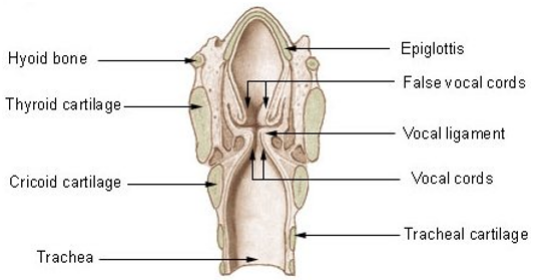


Figura 2.2: Secção coronal da laringe e parte superior da traquéia

2.1.2 Músculos e Cartilagens

Os músculos e cartilagens atuam diretamente no processo de abdução e adução das pregas vocais. Estas estão localizadas dentro da laringe e devido à dinâmica das cartilagens e dos músculos, podem executar os movimentos citados de forma a produzir som.

Cartilagens da Laringe

A Figura 2.2, mostra uma secção da laringe, detalhando as cartilagens presentes. De maneira sucinta, estas cartilagens servem como base de interconexão para os músculos intrínsecos ao redor da laringe. Dentre as cartilagens acima, a epiglote é responsável por vedar as vias respiratórias movimentando-se sobre a entrada das mesmas. O resto das cartilagens garantem a mobilidade da laringe em conjunto com outras estruturas como por exemplo o sternum.

2.2 Propriedades Físicas

2.2.1 Lei Bernouli

Energia potencial e energica cinética em fluídos se mantém a mesma porém em proporções diferentes (15):

$$P+\frac{\rho * v^2}{2} = ConstanteSendo : \rho = Densidadedofluido P = Pressonodutoondeofludoseencontrav = v$$

2.2.2 Critérios para Oscilação

Alguns critérios devem ser atendidos para que um determinador 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 (16), 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 Tensao

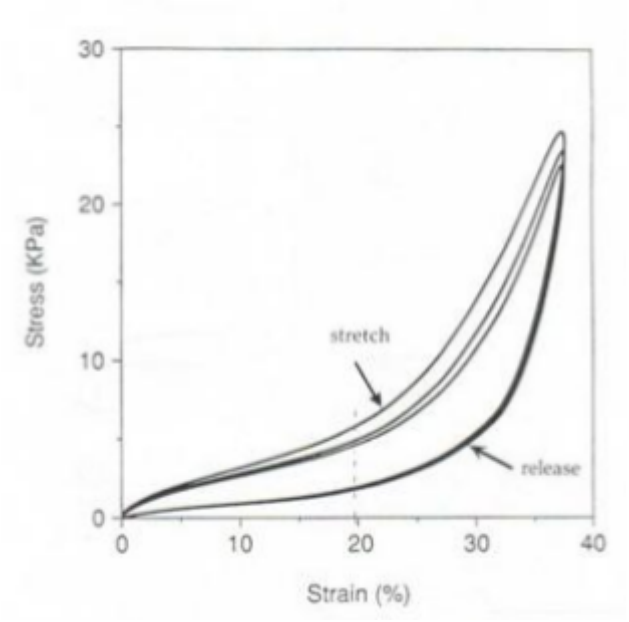
Conceito de Tensão For por unidade de ar

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

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 materias 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 Reflexao de Som

Um fenômeno ligado a rigidez e amortecimento entre um meio e outro.(9) Ondas quando tentam penetrar em um segundo meio, sendo o segundo meio rígido, as partículas primeiro meio se aglomeram tentando passar porém falham, seu acúmulo de particulas geram pressão que acabam criando uma outra onda no primeiro meio decorrente da primeira onda.(6)

O mesmo ocorre com o meio 2 sendo totalmente não rígido e o primeiro meio sendo bem rígido, Exaurindo excesso de particulas 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).

2.2.8 Fluxo de Ar na Glote

Fluxo de ar na Glote:

Como descrito no artigo de Elias temos informações como descobrimos a pressão via fluxo de ar(2)

$$U_g = + - (\frac{-a_m}{A_*} + [(a_m)^2 + -(\frac{4K_t}{C^2\rho})](P_s^+ - P_i^-))$$

A^* = Área efetiva computada pelas areas $A_i e A_s$Área efetiva computada pelas areas $A_i e A_s$Área efetiva com

$$\rho = \text{Densidade do Ar}$$

$$c = \text{Velocidade do som}$$

$$P_s e P_i = \text{Pressão de incidencia na entra e saída da glote}$$

Uma vez descoberto o fluxo, as pressões de reflexão $P_e^s P_i^+ podemserencontradascomaseguinteequao$:

$$P_s^- = P_s^+ - (\rho c / A_s) U_g$$

$$P_i^+ = P_i^- - (\rho c / A_i) U_g$$

Quando ocorre o fechamento da glote então alguns parametros assumem valores conhecidos. a(t) = 0 , $P_g = \frac{P_s - P_i}{2}$ e $U_g = 0$

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

$$P = (1 - \frac{a2}{a1}) * (Ps - Pi) + Pi$$

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 subglotica
- 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 (5) 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 cilindricos os quais se propagam uma onda acustica.Superglótica não leva em conta o trato vocal.Perdas e viscosidade são considerada.

Requisitos Funcionais

O Synpath é consistido também dos seguintes requisitos funcionais, os requistos funcionais são as funcionalidades que o sistema executará(4)

- 1 - Validação do Parâmetros passados pelo Usuário, se condizem com restrições do programa.
- 2 - Plotar um gráco inicial do trato vocal de acordo com os parâmetros do usuário.
- 3 - Plotar três grácos referentes as propriedades da voz simuladas com os parâmetros fornecidos pelo usuário. – O primeiro gráco refere-se às posições adotadas pelas cordas vocais, a área da glótis, ao uxo de ar nessa área e às características desse uxo. – O segundo gráco refere-se às características do som gerado pela simulação física do aparato fonador pelo programa. – O terceiro gráco refere-se ao espectro de frequência do som gerado e do uxo da glótis.
- 4 - Gerar um arquivo de texo com as características de voz gerada, frequencia,amplitude, ruído da voz, entre outros
- 5 Gerar um arquivo de som de voz simulada.

Requisito Não-Funcionais

O Synpath consiste também dos seguintes requisitos não funcionais, requisitos não funcionais são requisitos são parametros de qualidade, requisitos que limitam as funcionalidades do sistema(4).

- 1 - O Sistema deve produzir os gráficos que os requisitos funcionais delimitaram em um intervalo de 1(um) minuto.
- 2 - Após gerar os gráficos e os exibi-los o arquivo texto e o arquivo de audio deverão ser exibidos
- 3 - Para que o sistema esteja funcional é necessário ter instalado os pacotes: MatPlotLib NumPY
- 4 - O sistema deve ser executado em plataformas de um sistema operacionais como Windows, Linux ou MacOs, Versões recentes de acordo com a data desse documento.

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 (3).Após isso as HMM's são treinados através dos dados de treinamento usando o algoritmo EM.(7).

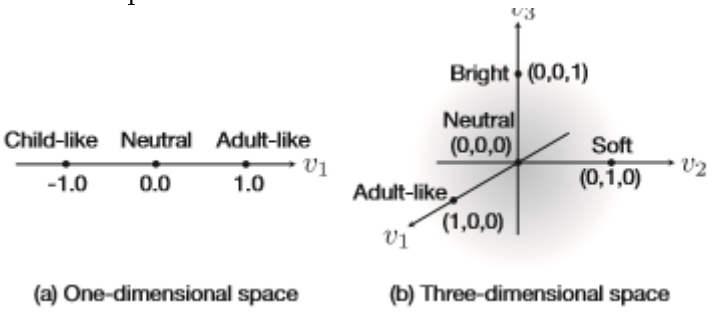
Após isso ocorre a fase de sintesis, usa-se outra partitura para ser convertida em "labels"dependentes de contexto e estima-se quais "labels"preprocessadas são correspondentes.(12)

2.3.4 MHRSMM

Uma variação HSMM. Modelo de multipla regressão HSMM. Parametros importantes são i e mi dos outputs pdfs

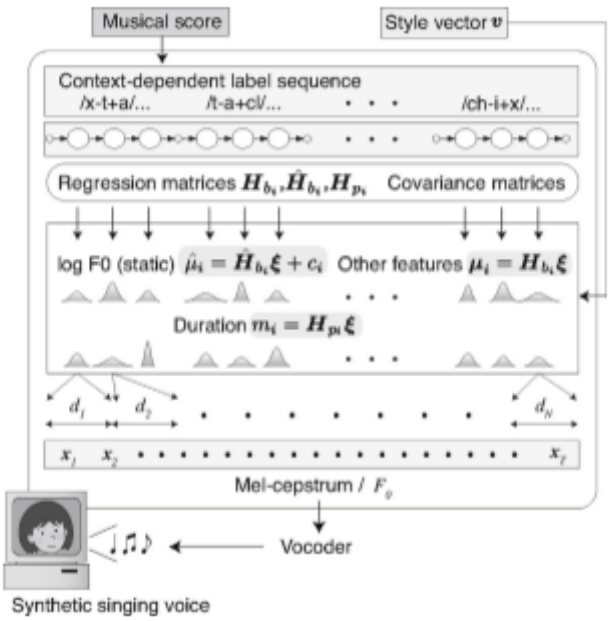
$$\mu_i = H_{bi}\xi$$
$$m_i = H_{pi}\xi$$
$$\xi = [1, v_1, v_2, ..., v_L]^T$$
$$\xi = [1, v^T]^T$$

Onde L é a dimensão do vetor de estilo e vi é a intensidade do enésimo estilo de canto. Um exemplo de um vetor de estilos de canto de tamanho L = 2 e L = 3.



(a) One-dimensional space (b) Three-dimensional space

Controle do Sintentisador de voz cantada baseado em MRHSMM



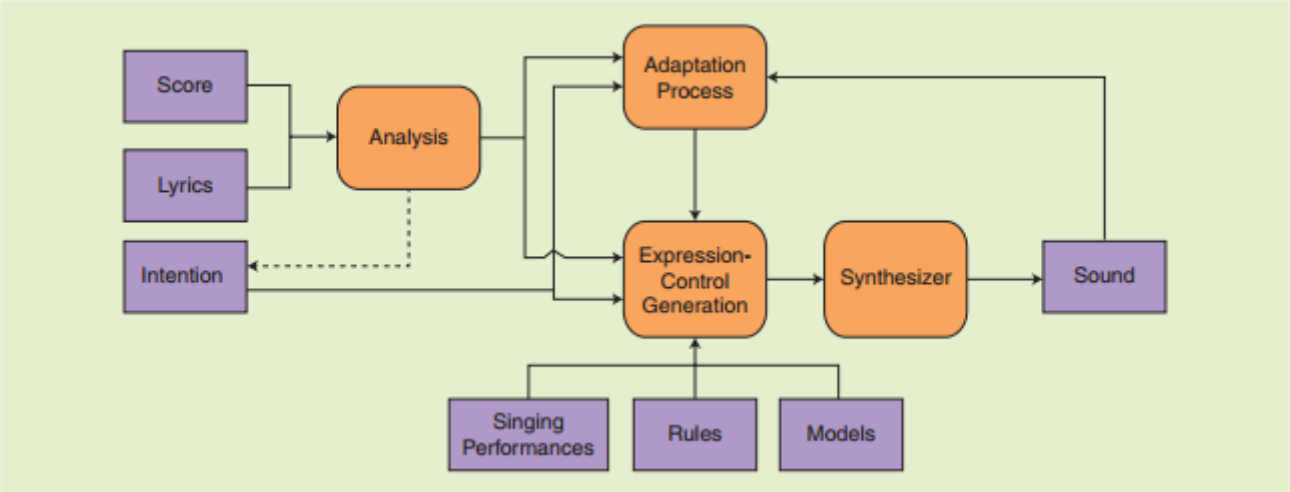
Durante a fase de síntese o usuário do programa adiciona vetores de estilos de acorodo com a intenção e a expressividade pretendida. Parametros de output como duração são gerados pelos vetores de estilos dados e matrizes de regressões treinadas usando MRHSMMs

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

MRHSMM possui uma dificuldade de gerar contorno F0 que acompanhe o contexto de mudança de altura das notas o author TAKASHI NOSE, propõe um treinamento de HSMM e HMM nos parametros

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.(17)

Expressão

Expressão músical é um conceito intuitivo porém dificil de se definir. A expressão é chave na percepção da qualidade e naturalidade musical. No caso da voz cantada implica-se usar vários outros parametro além de frequencia e amplitude. Psicológicamente contorno do timbre, vibrato, tremolo, timing fonético.(17)

2.3.6 Envoltoria F0

Envoltórias F0 são usadas para expressar informação linguistica, para-linguística e não-linguistica.(13)

As Envoltória F0 apresentam três (3) características importantes que fazem diferenciar uma voz falada a uma voz cantada.(14)

- 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 muita flutuações f0 que são apenas observadas em apenas vozes cantadas

2.3.7 Sintese de Voz em Mandarím

Utiliza-se a técnica HNM para a sintese da voz cantada em mandarím. 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.(11)

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.(10)

Sistema auditivo diferencia-se entre espécies em jeitos interessantes. Por exemplo algumas espécies tem caracteristicas diferentes relacionadas aos sinais vocais mais utilizados por ela mesma.(10)

2.4.2 Intensidade

Como a frequência de um estímulo a intensidade dele é processado e codificado subcortical nos dois lados do cérebro em todos os níveis no cerébro.(10)

2.5 Voz e Propriedades Linguísticas

Uma divisão importante de acordo com Flanagen, são as letras separados em classificações Vogais e Consoantes que se associam a um movimento do trato vocal correspondente (8).

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 articação normal, mantém-se relativamente estável.Há uma opção de contribuição das cavidades nasais, uma cobertura porém é negligenciável.Baseado nessas caracteristicas é divididos todas as consonantes. A tabela abaixo explica:

2.5.2 Consoantes

Sons produzidos com constrições 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. 5

[2] Elias Amadeu de Souza. Simulação computacional de uma fenda glotal, 2014. 5

[3] 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. 7

[4] Pierre Bourque & Richard E.(Dick) Farlay. *Guide to the Software Engineering Body of Knowledge Version 3.0.* IEEE Computer Society, 2004. 7

[5] Jean Schoentgen Samia Fraj and Francis Grenez. *Development and perceptual assessment of a synthesizer of disordered voices.* Journal of Acoustical Society of America, 2012. 6

[6] Henry Gray. *Anatomy of the Human Body.* IEEE Tran. Audio Eletroacoust. 5

[7] 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. 7

[8] James L. Flanagan Jont B. Allen Mark A. Hasegawa-Johnson. *Speech Analysis Synthesis and Perception.* 2008. 10

[9] 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. 5

[10] Jody Kreiman and Diana Van Lancker Sidtis. Foundations of voice studies, 2011. 2, 10

[11] 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. 10

[12] 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. 8

[13] 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 GenerationInternational Conference on Speech Prosody. 9

[14] 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 GenerationInternational Conference on Speech Prosody. 10

[15] BradH Story. *Tubetalker*. Dept. of Speech, Language, and Hearing, Tucson, AZ, 2010. 4

[16] Ingo Titze. *Principles of Voice Production*. Prentice Hall, New Jersey 07632, 1994. 4

[17] 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. 9