Morfologia do trato vocal por imagens de ressonância magnética em tempo real

Rafael de Assunção Sampaio

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

Programa: Mestrado em Ciência da Computação

Orientador: Prof. Dr. Marcel Parolin Jackowski

São Paulo, dezembro de 2016

Morfologia do trato vocal por imagens de ressonância magnética em tempo real

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 07/12/2016. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Marcel Parolin Jackowski (orientador) - IME-USP
- Prof. Dr. Paulo André Vecchiato de Miranda - IME-USP
- Prof[a]. Dr[a]. Fátima de Lourdes dos Santos Nunes Marques - EACH-USP

Agradecimentos

Em primeiro lugar, quero agradecer à Universidade de São Paulo, que me acolhe desde a Graduação. Sou grato pelas oportunidades acadêmicas e pelo crescimento intelectual que me foram propiciados.

Meu agradecimento, especialmente, ao Instituto de Matemática e Estatística, pelo ambiente amistoso e colaborativo construído pela maioria dos docentes, alunos e funcionários.

Aos professores e colegas das edições de 2014 das disciplinas de MAC5918 e MAC5768 do IME-USP, ministradas pelos Prof. Marcel Jackowski e Prof. Roberto Hirata Jr., pelas ideias que levaram à investigação de várias das abordagens presentes neste trabalho. Ao Prof. Paulo Miranda pelo interesse e sugestões de técnicas a serem pesquisadas.

Ao meu orientador, o Prof. Marcel Jackowski, pelo tempo, dedicação e paciência. Muito obrigado pelo incentivo e ajuda ao longo desta pesquisa!

Por fim, à minha querida família – Vera, Veronica, Simone, Lara, Teresa –, aos meus estimados amigos que acompanham a evolução deste trabalho, muitíssimo obrigado por existirem em minha vida.

Rafael Sampaio

Resumo

SAMPAIO, R. A. **Morfologia do trato vocal por imagens de ressonância magnética em tempo real**. 2016. 51 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

A técnica de imagens de ressonância magnética em tempo real (RM-TR) possibilitou, de forma inédita, a observação da dinâmica dos processos ocultos de articulação. A natureza não invasiva das imagens por RM-TR combinada com o seu poder de discriminação anatômica fez que esta técnica se tornasse a referência na captura da configuração do trato vocal durante a produção da fala. Este avanço, no entanto, trouxe também alguns desafios, entre eles a extração automática de contornos do trato vocal e a análise de sua forma a partir de tais imagens. Esta dissertação traz técnicas automatizadas de segmentação do trato vocal e a identificação de suas estruturas articulatórias. Em especial, a identificação de tais estruturas são vitais para a modelagem da síntese articulatória. A metodologia se baseia em curvas de nível (*level sets*) para o delineamento das bordas do trato vocal. A variação da forma do trato vocal e de suas estruturas foi investigada para diferentes *córpora*, com vistas ao paralelismo entre a expressão dos fonemas e o comportamento das formas anatômicas. Estas foram rotuladas a partir de invariantes da forma basal, cuja evolução reproduziu a classificação. A metodologia resultante poderá ser utilizada em aplicações inovadoras, como na criação de sistemas para a supressão de sotaque, auxílio à produção da fala em pacientes laringectomizados e terapia de crianças com apraxia da fala.

Palavras-chave: curvas de nível, contornos ativos, morfologia, trato vocal, ressonância magnética, tempo real, RM-TR.

Abstract

SAMPAIO, R. A. **Vocal tract morphology using real-time magnetic resonance imaging**. 2016. 51 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Real-time Magnetic Resonance Imaging (rtMRI) leads to the dynamic observation of hidden processes of articulation in an unprecedented way. The non-invasive nature of the images combined with its power of anatomical discrimination made this technique the reference in acquiring the vocal tract configuration during speech production. However, this development also brought some challenges that include the extraction and shape analysis of the vocal tract contours from such images automatically. This work brings automated techniques for segmentation of the vocal tract and identification of articulatory structures. In particular, the identification of such structures is vital for modeling articulatory synthesis. This methodology is based on level set method to outline the vocal tract shape. The change in shape of the vocal tract and its structures was investigated for different corpora in order to parallelism between the expression of phonemes and the behavior of the anatomical shapes. These shapes were labeled from invariants of basal form, whose evolution brought classification of regions of interest. The methodology resulting from this work may be employed in innovative medical applications, such as accent-suppression systems, speech production for laryngectomized patients, and therapy for children suffering from speech apraxia.

Keywords: level set, active contours, morphology, vocal tract, magnetic resonance imaging, real time, rtMRI.

Sumário

Lista de Abreviaturas

RM	Ressonância Magnética
TR	Tempo Real
AL	Abertura Labial
CL	Constrição da Língua
AVP	Abertura Véu Palatino
DICOM	*Digital Imaging and Communications in Medicine*
LSF	Função de Curva de Nível (*Level Set Function*)
INCA	Instituto Nacional de Câncer José Alencar Gomes da Silva
OMS	Organização Mundial de Saúde
VP	(Resultados) Verdadeiros Positivos
VN	(Resultados) Verdadeiros Negativos
FP	(Resultados) Falsos Positivos
FN	(Resultados) Falsos Negativos
TVP	Taxa de Verdadeiros Positivos
TVN	Taxa de Verdadeiros Negativos
TFP	Taxa de Falsos positivos
TFN	Taxa de Falsos Negativos

Lista de Símbolos

$I * K$	Convolução de I pela máscara K
G_σ	Máscara gaussiana com desvio-padrão σ
g	Função indicadora de bordas
δ	Função δ de Dirac
H	Função H de Heaviside
ϕ	Curva de nível
$C(s,t)$	Curva paramétrica (espaço s e tempo t)
$\mathcal{F}$	Função que controla a velocidade do movimento da curva
$\mathcal{N} = -\frac{\nabla\phi}{\lvert\nabla\phi\rvert}$	Vetor normal e interno à curva
$\mathcal{E}(\phi)$	Funcional de energia de ϕ
$\mathcal{E}_{ext}(\phi)$	Energia externa a ϕ
p	Função de potencial
$\mathcal{R}_p(\phi)$	Termo de regularização de ϕ com potencial p
$\nabla\phi$	Gradiente de ϕ
$\nabla^2\phi$	Laplaciano de ϕ
div $\mathbf{F}$	Divergente do campo vetorial $\mathbf{F}$
J	Medida Jaccard
D	Medida Dice
H	Distância Hausdorff

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Vox nihil aliud quam ictus aer.
(A voz não é nada além de ar vibrante.)

– Séneca (filósofo, 4 a.C.-65)

A síntese articulatória procura produzir a fala através de modelos do trato vocal e de seus processos articulatórios. Isso é feito por meio da modelagem da forma do trato vocal e do fluxo do ar que faz vibrar as cordas vocais. A Fig. 1.1 ilustra os contornos de interesse das componentes do aparelho fonador: laringe, epiglote, língua, lábios, parede da faringe, glote, véu palatino e palato duro (céu da boca). De acordo com Bresch *et al.* (2008), esses oito componentes anatômicos (com exceção do palato duro) são chamados articuladores da fala, os quais são controlados durante a produção da fala. O conhecimento a respeito da posição e movimentos desses articuladores é fundamental em estudos da produção da fala.

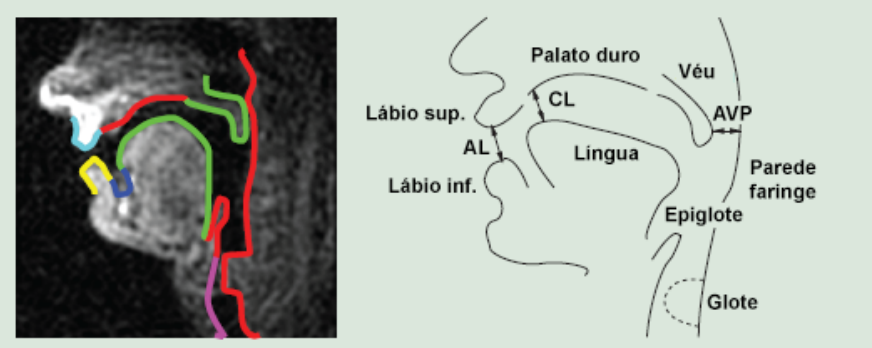


Figura 1.1: *Corte sagital de imagem RM-TR sobreposta com contornos dos articuladores da fala. Os articuladores e variáveis do trato vocal, como abertura labial (AL), grau de constrição da ponta da língua (CL), e abertura do véu palatino (AVP) são ilustrados. Adaptado de Bresch e Narayanan (2009).*

Várias técnicas de aquisição de imagens têm sido utilizadas para observar os processos intrínsecos da articulação. Entre elas, podemos citar o ultrassom (Whalen *et al.* (2005)), raios X (Fontecave e Berthommier (2006)), a articulometria eletromagnética (Perkell *et al.* (1992)) e a ressonância magnética (Alvey *et al.* (2008); Badin *et al.* (1998); Baer *et al.* (1991); Demolin *et al.* (2002)). A natureza não invasiva das imagens de ressonância magnética (RM) combinada com o seu poder de discriminação anatômica fez que esta técnica se tornasse a referência na captura da configuração do trato vocal durante a produção da fala. Desde o primeiro estudo proposto por Demolin *et al.* (2002), muitos estudos foram conduzidos utilizando RM, entre os quais a produção das vogais; a produção de consoantes; e em diferentes línguas, tais como francês, alemão, japonês, português europeu e português brasileiro (Gregio (2006); Martins (2011)). Avanços na técnica de RM fizeram que aquisições pudessem ser realizadas em tempo real (RM-TR), em que um grande número de imagens são geradas de forma a capturar a dinâmica da articulação (Narayanan *et al.* (2004)).

Este avanço, no entanto, trouxe também vários desafios, entre eles a extração automática de contornos do trato vocal a partir de tais imagens. Enquanto bordas entre diferentes tecidos podem ser delineadas manualmente com grande acurácia (Demolin *et al.* (1996); Lecuit (1992); Stone *et al.* (2001)), tais abordagens são laboriosas, sujeitas a inconsistências entre quadros, e tornam-se inviáveis em grandes sequências de imagens capturadas em tempo real. Como consequência, técnicas automatizadas de segmentação do trato vocal a partir de RM-TR são cruciais para sua eficiente utilização em modelos de síntese articulatória.

Resultados recentes para a extração do trato vocal dependem da interação humana, tornando-se apenas técnicas semiautomáticas, seja pela base de treinamento das segmentações (Bresch e Narayanan (2009)), seja pela identificação de pontos-chave para a estimação da curva da cavidade aérea (Kim *et al.* (2014)). A atribuição automática de tais pontos foi explorada por Raeesy *et al.* (2013), com a construção de uma base de treinamento analisada por componentes principais, dependente de várias segmentações manuais de entrada também.

1.1 Contexto de aplicações

Comunicar-se em idiomas diferentes aproxima culturas e promove uma forma eficiente de compartilhar conhecimento. No entanto, quanto à comunicação oral, a eficiência é um fator que depende da habilidade do falante em conseguir expressar as suas ideias, de forma clara e inteligível, para o seu público-alvo. O aprendizado de um segundo idioma, por vezes, é imperfeito, isto é, o falante é capaz de aprender, mas não reproduz com exatidão todos os fonemas, sofrendo com atitudes discriminatórias e com estereótipos negativos.

Cenários semelhantes se materializam por meio de patologias, como ocorre com pessoas laringectomizadas. Tais pessoas possuem uma alteração dos mecanismos de condução do ar até os pulmões, devido à remoção parcial ou total da laringe para tratamento de tumores. A laringectomia acarreta a perda da voz laríngea, podendo ocorrer a reabilitação através da voz esofágica, da voz tráqueo-esofágica, ou uso de próteses fonatórias. Esse tratamento é conduzido por fonoaudiólogo, para que um treinamento adequado permita a recuperação da fala (INCA (2015)).

Outra patologia é a apraxia da fala em crianças. De acordo com a *American Speech-Language-Hearing Association*, trata-se de um distúrbio neurológico que afeta a produção motora dos sons da fala. A precisão e a consistência dos movimentos necessários para produzir os sons da fala são alteradas, na ausência de déficits neuromusculares. O tratamento também é conduzido por fonoaudiólogo e os resultados podem ser de longo prazo (Ziegler *et al.* (2012)).

Todos os casos mencionados requerem a compreensão do trato vocal do indivíduo. Para tanto é essencial identificar, automaticamente, a deformação do trato vocal e de suas estruturas. Tal delineamento levaria, num segundo momento, à tentativa de suprimir tais deficiências ao lado de profissionais da área de fonoaudiologia.

1.2 Objetivos

Tendo em vista que a técnica de RM-TR oferece grandes vantagens em relação a outras técnicas na identificação da dinamicidade das estruturas do trato vocal, investigamos metodologias para a segmentação dessas imagens, cujos contornos finais poderão subsidiar outros trabalhos, combinando-se dados acústicos e transcrições fonéticas. Os objetivos específicos são:

1. Extração do trato vocal de imagens RM-TR de forma automática, utilizando similaridades entre os seus contornos e o conhecimento *a priori* da forma do trato vocal;

2. Identificação automática das estruturas articulatórias: lábios, língua, palato duro, véu palatino, parede da faringe, glote e epiglote.

1.3 Contribuições

Com a consolidação dos objetivos mencionados, as principais contribuições deste trabalho são as seguintes:

1. Metodologia e implementação de segmentação automática de imagens de ressonância magnética em tempo real do trato vocal.

   1.1 Outros métodos na literatura resolvem o problema para um articulador específico, utilizando hipóteses estruturais ou a interação com o usuário ao longo do processo.

   1.2. A não invasibilidade do método, em particular, preserva as características naturais da fala, sendo robusto à ausência de um ou mais articuladores e a diferentes contextos fonéticos dos idiomas.

2. Classificação das estruturas articulatórias do trato vocal, a partir da definição de pontos invariantes, que possibilitará a compreensão da dinâmica destas na busca de terapias.

3. Disponibilidade dos dados adquiridos para utilização em pesquisas, bem como o fornecimento do protocolo de aquisição das imagens a pesquisadores interessados.

1.4 Organização do Trabalho

No Capítulo 2, apresentamos os conceitos estudados para a segmentação do trato vocal. No Capítulo 3, elencamos os trabalhos relacionados. No Capítulo 4, desenvolvemos a metodologia a que se propõe esta dissertação. No Capítulo 5, exibimos os resultados de segmentação do trato vocal. No Capítulo 6, apresentamos as conclusões obtidas neste trabalho, bem como as vantagens e desvantagens da metodologia proposta.

Capítulo 2

Conceitos

Ofereça a todo homem o teu ouvido, mas a poucos a tua voz.

– Shakespeare (escritor, 1564-1616)

Esta pesquisa explora as formulações de contornos ativos e curvas de nível, para abordar o delineamento do trato vocal. Apresentaremos os conceitos fundamentais, para avançar na construção da metodologia que será proposta. Fará parte da metodologia o pré-processamento das imagens de RM-TR, para o qual será especialmente relevante o filtro gaussiano, resultado da convolução da imagem com uma máscara específica. Além disso, definiremos a função δ de Dirac e a função H de Heaviside, bem como as suas aproximações, o divergente de um operador e, brevemente, as diferenças finitas para as derivadas parciais de uma curva de nível. Gonzalez e Woods (2006) é uma referência para várias das definições.

2.1 Imagem digital

Uma imagem pode ser definida como uma função bidimensional $f(x, y)$, em que x e y são coordenadas espaciais, e a amplitude de f em qualquer par (x, y) é chamada de intensidade da imagem naquele ponto. Quando x, y e f têm valores finitos, quantidades discretas, a imagem é chamada de digital. A estrutura de dados matricial é a mais frequente para representá-la.

2.2 Imagens de Ressonância Magnética

Segundo McRobbie *et al.* (2003), a RM se baseia, fundamentalmente, num intenso campo magnético ao redor da região de interesse. Dessa forma, o paciente posicionado dentro da máquina de RM tem os prótons contidos em suas moléculas alinhados na mesma direção. Ondas de rádio são, então, emitidas pela máquina numa radiofrequência específica[1], deslocando os prótons de seu alinhamento. Quando as ondas cessam, os prótons se realinham e devolvem a energia absorvida em forma de ondas enquanto retornam para seu estado original.

Os prótons se realinham em diferentes velocidades em cada tipo de tecido, o que leva ao contraste presente na imagem. O retorno ao estado original de equilíbrio ocorre segundo uma curva exponencial de parâmetro T_1, também chamado de *tempo de relaxamento* T_1.

Para obter uma imagem de RM neste experimento, o paciente veste uma bobina específica na cabeça (semelhante a um elmo), é colocado sobre uma maca e orientado a manter o corpo imóvel. Por deslizamento, é introduzido no túnel do aparelho, sendo possível ouvir instruções da mesa de controle antes que o procedimento se inicie. A possibilidade de sedação, para pacientes claustrofóbicos ou crianças, é inconveniente, pois ele deverá falar enquanto estiver no aparelho; em vez disso, estratégias psicoterápicas de dessensibilização podem ser aplicadas previamente.

[1] Conhecida por frequência de Larmor, homenagem ao físico britânico Joseph Larmor (1857 - 1942).

As imagens denotadas por "em tempo real" se referem à rápida e contínua aquisição de imagens de RM, cujo elemento de interesse é um processo dinâmico – neste caso, o processo articulatório do trato vocal. Um equilíbrio torna-se necessário entre a resolução espacial e a resolução temporal – quando uma delas é elevada, a outra é degradada.

Cabe destacar que a RM utiliza radiação não-ionizante, o que não traz prejuízos à saúde na tomada recorrente de imagens. Mas cuidados devem ser tomados em relação a pacientes que possuam implantes de metal, tatuagens, sensibilidade auditiva, estimulação nervosa, entre outros, cumprindo-se o protocolo de segurança à saúde, conforme definido pela *International Electrotechnical Commission.*

2.3 DICOM e metadados

DICOM é um padrão internacional para gerenciamento, armazenamento, impressão e transmissão de imagens médicas (ISO 12052).

As bases de dados que adotam esse padrão reúnem, de forma íntegra, imagens e uma coleção de atributos, conhecidos por *metadados*, que se referem ao paciente, ao médico, ao aparelho em que a imagem foi obtida, entre outros (NEMA (2016)).

2.4 Robustez e Consistência

A definição formal para algoritmo robusto em Visão Computacional teve origem, segundo Vacavant (2016), nas pesquisas de Peter Meer. A definição não é construtiva, o que dificulta a sua aplicação diretamente: "*Robustez em visão computacional não pode ser atingida sem que se tenha acesso a um valor razoavelmente correto da escala do ruído*". Em nosso contexto, as transições ao longo do processo articulatório do trato vocal compreendem o ruído. O método avaliado, nesse sentido, será mais *robusto* quanto mais adaptativo às transformações do trato vocal.

Berry (2007) define consistência como "*necessária para tornar válida a combinação ou comparação de resultados a partir de dados adquiridos em momentos diferentes, de maneira que o mesmo procedimento analítico possa ser aplicado*". A consistência, em nosso contexto, está relacionada à reprodutibilidade do método de segmentação a partir de novos dados: é esperado que processos pouco dependentes da interação humana apresentem maior consistência.

2.5 Convolução de um operador

Sejam I a matriz que representa uma imagem, K uma máscara (de dimensões $2k+1$ por $2k+1$, com k inteiro positivo) e R a matriz resultante. Define-se a convolução de I pela máscara K ou, simplesmente, $I * K$, por

$$R(i,j) = \sum_{u=-k}^{k} \sum_{v=-k}^{k} K(u,v) I(i-u, j-v).$$

2.6 Filtro gaussiano

Define-se a função gaussiana centrada em $(0,0)$ com desvio-padrão σ ou, simplesmente, G_σ, por

$$G_\sigma(x,y) = e^{-\frac{x^2+y^2}{2\sigma^2}}.$$

O filtro é isotrópico, isto é, apresenta simetria circular e σ determina a intensidade de suavização a ser aplicada.

A imagem é armazenada como uma coleção discreta de pixels. É necessário produzir também uma aproximação discreta para aplicar a função gaussiana, isto é, uma máscara que será utilizada

na convolução. A função gaussiana nunca se anula, logo poderia requerer uma máscara de tamanho infinito. Isso não é necessário, no entanto, pois ela se aproxima de zero à medida que se afasta da média (centro) por mais de três desvios-padrões. Ademais, uma implementação mais eficiente pode considerar que $G(x, y) = G(x)G(y)$, reduzindo a convolução ao caso unidimensional.

Este filtro é utilizado para borrar as imagens, removendo ruído e detalhes.

2.7 Função indicadora de bordas

Seja I a matriz que denota a imagem, definimos g um indicador de borda por

$$g = \frac{1}{1 + |\nabla G_\sigma * I|^2}.$$

Essa função admitirá, em geral, valores menores nas bordas da região de interesse.

2.8 Função δ de Dirac

Caracterizamos a função δ de Dirac como uma função na reta real que é zero para todos os pontos de seu domínio, exceto na origem, onde é infinita.

$$\delta(x) = \begin{cases} +\infty & \text{se } x = 0 \\ 0 & \text{se } x \neq 0 \end{cases}$$

Além disso, ela deve satisfazer à condição:

$$\int_{-\infty}^{\infty} \delta(x)dx = 1.$$

É relevante explicitar que esta não é uma definição matemática, pois não existe nenhuma função nos reais que satisfaça a essa caracterização (Dirac (1958)). No entanto, é suficiente para a construção da aproximação utilizada por Li *et al.* (2010).

$$\delta_\epsilon(x) = \begin{cases} \frac{1}{2\epsilon}(1 + \cos(\frac{\pi x}{\epsilon})) & \text{se } |x| \leq \epsilon \\ 0 & \text{se } |x| > \epsilon \end{cases}$$

No processamento das imagens, x é ocupado pela função curva de nível (definida adiante). O parâmetro ϵ adotado é 1,5.

2.9 Função H de Heaviside

Observamos que a função δ de Dirac é a derivada da função H de Heaviside. A aproximação utilizada por Li *et al.* (2010) é

$$H_\epsilon(x) = \begin{cases} \frac{1}{2}(1 + \frac{x}{\epsilon} + \frac{1}{\pi}\sin(\frac{\pi x}{\epsilon})) & \text{se } |x| \leq \epsilon \\ 1 & \text{se } x > \epsilon \\ 0 & \text{se } x < -\epsilon \end{cases}$$

No processamento das imagens, x também é ocupado pela função curva de nível. Essa definição é consistente com a aproximação da função δ de Dirac.

2.10 Contornos Ativos e Curvas de Nível

Os matemáticos Osher e Sethian (1988) disseminaram a metodologia de representação de contornos que utilizam curvas de nível zero de uma função de dimensão imediatamente superior à do espaço em que a curva original está descrita. Essa função de dimensão superior é conhecida por

função de curva de nível (*level set function*, LSF). O livro de Osher e Fedkiw (2003) reúne grande parte do conhecimento nesta área de pesquisa.

A evolução de uma curva de nível pode ser deduzida a partir de um modelo de contorno ativo (também conhecido por *snakes*, Kass *et al.* (1988)). Para tanto, define-se uma curva paramétrica dinâmica $C(s,t) : [0,1] \times [0,\infty] \to \mathbb{R}^2$, cujo parâmetro espacial é s e o temporal é t. A evolução da curva pode ser expressa por

$$\frac{\partial C(s,t)}{\partial t} = \mathcal{F}\mathcal{N}$$

em que $\mathcal{F}$ é a função que controla a velocidade do movimento do contorno, $\mathcal{N}$ é o vetor normal e interno à curva, dado por $\mathcal{N} = -\frac{\nabla\phi}{|\nabla\phi|}$. Considerando essa curva $C(s,t)$ como curva de nível zero de uma $\phi(x,y,t) : \mathbb{R}^2 \times [0,\infty] \to \mathbb{R}$, a equação diferencial parcial que define a LSF é obtida diretamente derivando-se ϕ em relação ao tempo:

$$\frac{\partial \phi}{\partial t} = \mathcal{F}\,|\nabla\phi|\,.$$

Este método apresenta vantagens estruturais, por permitir que mudanças topológicas sejam capturadas ao longo da evolução do contorno. As curvas de nível podem ser aplicadas diretamente ao plano cartesiano, sem necessidade de parametrização. No entanto, irregularidades numéricas podem ser encontradas ao longo da evolução. A função $\mathcal{F}$ pode depender de fatores internos, como curvatura, e externos, como gradiente da imagem. No Capítulo 4, estudaremos a teoria de curvas de nível desenvolvida no artigo de Li *et al.* (2010), que estende o conceito para uma formulação variacional.

2.11 Divergente de um campo vetorial

Seja $\mathbf{F}$ um campo vetorial continuamente diferenciável (classe C^1), definimos

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}.$$

Por exemplo, tomemos $\mathbf{F}(x,y,t) = \frac{\nabla\phi}{|\nabla\phi|}$. Temos que:

$$\text{div } \mathbf{F} = \nabla \cdot \frac{\nabla\phi}{|\nabla\phi|} = \frac{\nabla^2\phi}{|\nabla\phi|}.$$

Em particular, $\nabla^2\phi$ representa o laplaciano de ϕ e div $\mathbf{F}$ denota a curvatura média das curvas de nível de ϕ.

2.12 Derivadas parciais de curva de nível aproximadas por diferenças finitas

Seja $\phi(x,y,t)$ uma curva de nível temporal em t. As derivadas espaciais $\frac{\partial\phi}{\partial y}$ e $\frac{\partial\phi}{\partial y}$ podem ser aproximadas por diferenças centrais da seguinte forma:

$$\frac{\partial\phi}{\partial x}(x,y,t) \approx \frac{\phi(x+\Delta x,y,t) - \phi(x-\Delta x,y,t)}{2\Delta x}$$

e

$$\frac{\partial\phi}{\partial y}(x,y,t) \approx \frac{\phi(x,y+\Delta y,t_0) - \phi(x,y-\Delta y,t)}{2\Delta y}.$$

A derivada temporal $\frac{\partial\phi}{\partial t}$ pode ser aproximada pela diferença posterior:

$$\frac{\partial\phi}{\partial t}(x,y,t) \approx \frac{\phi(x,y,t+\Delta t) - \phi(x,y,t)}{\Delta t}.$$

Tais aproximações serão especialmente úteis na implementação do modelo de evolução das curvas de nível. A definição dos incrementos Δx, Δy, Δt será discutida no Capítulo 4.

Capítulo 3

Revisão Bibliográfica

"Se eu vi mais longe foi por estar de pé sobre ombros de gigantes."

– Isaac Newton (físico e matemático, 1643 - 1727)

Neste capítulo, expomos o contexto da área de pesquisa, mencionando os principais trabalhos relacionados.

3.1 Critérios de Busca

As técnicas de segmentação de imagens colocam-se como um assunto de relevante interesse no processamento de imagens. No entanto, estudos específicos do trato vocal são ainda incipientes.

Os critérios utilizados como crivo da seleção e posterior análise se pautaram por:

- Correlação do estudo encontrado com o trato vocal (por exemplo, a dinâmica da língua);
- Pesquisadores e orientados de universidades ou laboratórios atuantes nesta área;
- Recorrência de publicações sobre o assunto.

3.2 Trabalhos Relacionados

Os primeiros esforços na segmentação do trato vocal a partir de RM foram realizados em imagens de RM estáticas até a década de 90, porém, desde então, a grande maioria das soluções contemporâneas baseiam-se na utilização da técnica de RM-TR.

Avila-García *et al.* (2004) desenvolveram, em trabalhos anteriores, pesquisa sobre trato vocal a partir de imagens de ressonância magnética dinâmicas de Southampton, que consistem basicamente da gravação simultânea de imagem e som. No entanto, o ruído nessas imagens tornou difícil a extração das formas. Os pesquisadores, então, limitaram-se ao problema da extração da forma da língua, que é um articulador altamente deformável. Para tanto, combinaram modelos de contornos ativos com a transformada de Hough com vistas ao tracking da língua na sequência das imagens.

Bresch e Narayanan (2009) descrevem um método não supervisionado de segmentação de regiões de uma imagem utilizando a sua representação no domínio das frequências. O algoritmo visa a processar extensas sequências de RM-TR do trato vocal humano, utilizando um modelo anatômico sintético como informação *a priori*, cujo ajuste aos dados observados é feito por meio de otimizações. O objetivo dos autores é extrair o contorno e a posição dos articuladores do trato vocal durante a produção da fala.

Eryildirim e Berger (2011), apesar do título do artigo se referir ao trato vocal, partem de uma pesquisa anterior de Berger e refinam o modelo que utilizam para a segmentação da língua. A metodologia compreende a Análise de Componentes Principais a partir de *shape priors* (segmentações manuais da língua de um falante de referência). Com essa modelagem, adotam o modelo de

Chan-Vese (derivação do modelo de Mumford-Shah) para a otimização em busca do contorno de melhor ajuste para novas instâncias. Nesse sentido, a principal contribuição dos autores é propor uma identificação automática dos limites da estrutura articulatória (no caso específico, a língua) e realizar a validação do modelo de referência.

Vasconcelos *et al.* (2011) desenvolveram uma pesquisa baseada no estudo de fonemas do português europeu. Os autores construíram uma distribuição de formas a partir de 21 sons, que representariam as características principais da articulação do trato vocal. Essa informação *a priori* é utilizada em conjunto com modelos de contornos ativos para segmentar o trato vocal de novas imagens.

Raeesy *et al.* (2013) empregaram uma combinação de duas técnicas, que envolvem a localização automática de pontos anatômicos (Rueda e Udupa (2011)) a partir de uma base de treinamento e análise de componentes principais – o que representaria um aperfeiçoamento em relação ao trabalho de Vasconcelos *et al.* (2011) – e modelos deformáveis orientados (Liu e Udupa (2009)), para delinear as bordas do trato vocal em bases de dados potencialmente grandes.

Lammert *et al.* (2013) propuseram o uso de uma média das intensidades dos pixels numa dada região de interesse (no caso, a cavidade aérea), a fim de detectar a constrição do trato vocal na sequência de imagens.

Recentemente, Silva e Teixeira (2015) estendem a abordagem de Raeesy *et al.* (2013); Vasconcelos *et al.* (2011), utilizando dois modelos de aparência ativa para explícita e manualmente separar os fonemas orais dos nasais. Esse trabalho considera bases pequenas de treinamento e sustenta-se com a hipótese de baixa transição entre as imagens (i.e., resolução temporal maximizada).

Embora todos esses esforços tenham contribuído para a extração dos parâmetros do trato vocal a partir de RM-TR, essa área é ainda incipiente em soluções eficazes de segmentação, devido à grande variabilidade da forma da via aérea superior introduzida por sons distintos, à sua variabilidade entre diferentes falantes, à sua conectividade com outros canais, como a laringe e cavidade nasal, e à presença de ruído nas imagens.

Capítulo 4

Metodologia

Pois as palavras do ano findo pertencem à linguagem do ano findo,
as palavras do ano próximo outra voz aguardam.

– T. S. Eliot (escritor, 1888-1965)

Neste capítulo, caracterizamos as imagens com as quais trabalhamos e, então, apresentamos a nossa abordagem para cada um dos objetivos colocados. Esta combinará informações de baixo nível (intensidades dos *pixels*) com informações *a priori* de alto nível (forma e posicionamento dos articuladores da fala), visando a compensar a baixa resolução especial das imagens e a alta variabilidade das estruturas articulatórias. Finalmente, discutimos uma forma de validação dos resultados obtidos.

4.1 Imagens de RM-TR

As imagens de RM-TR utilizadas para o desenvolvimento e testes da metodologia de segmentação são de falantes nativos, que não apresentam deficiência de audição ou de fala, e são provenientes de:

1. Este projeto contou com uma parceria estabelecida com o Hospital São Paulo, da Universidade Federal de São Paulo (Unifesp), viabilizando aquisições de imagens de brasileiros. A tomada que obtivemos é de um homem que fala a língua portuguesa (nativo), com uma taxa de amostragem de 10 quadros por segundo; resolução espacial de 256×256 px^2 ($0{,}625 \times 0{,}625$ mm^2). Esse indivíduo falou três sentenças, marcadas com silêncio entre si: "Ela vê porco todo dia. Ela vê tigela todo dia. Ela vê carro todo dia." A série de imagens é constituída de 120 quadros.

2. Para testes e melhorias: um banco de dados de fala da USC-TIMIT[1], disponibilizado pela University of Southern California (USC). Esse banco de dados contém imagens de RM-TR de cinco homens e de cinco mulheres que falam a língua inglesa (nativos), com uma taxa de amostragem de 12,5 quadros por segundo; resolução espacial de 68×68 px^2 ($2{,}9 \times 2{,}9$ mm^2); e dados de articulometria eletromagnética de três desses indivíduos. Esses indivíduos, nas duas sessões de aquisição de imagens, repetiram o mesmo corpo de 460 sentenças foneticamente equilibradas.

[1] http://sail.usc.edu/span/usc-timit/

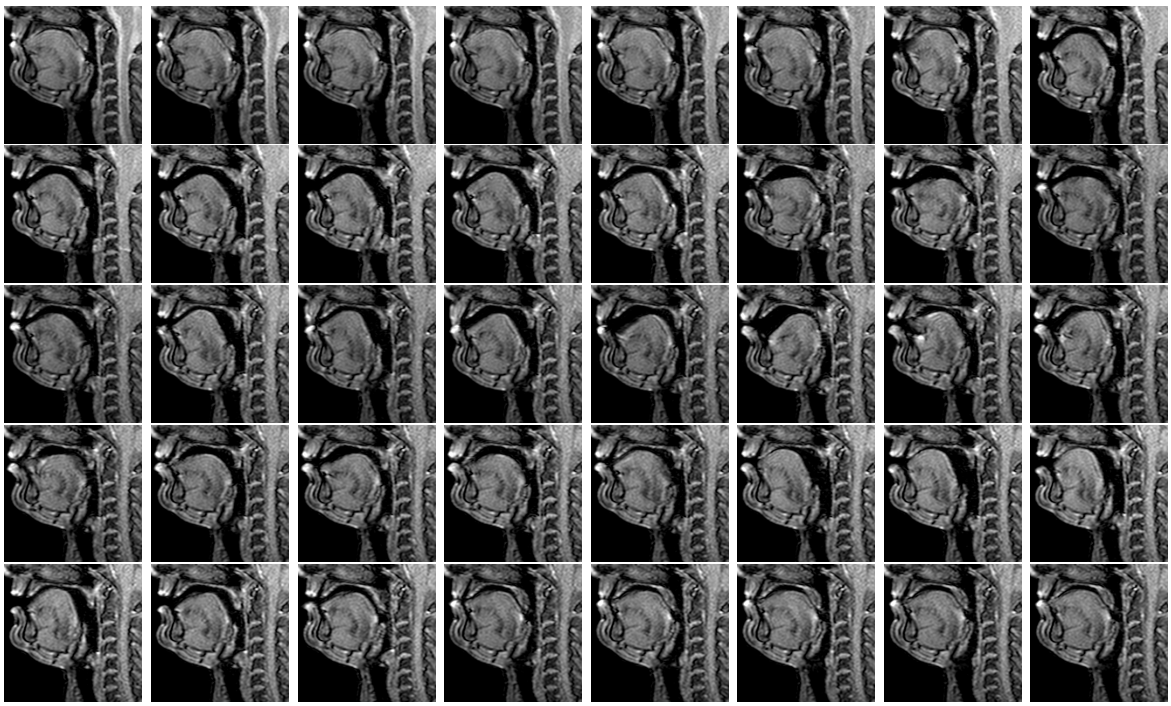


Figura 4.1: *Série de imagens que corresponde à sentença: "Ela vê porco todo dia".*

As imagens localizam o plano médio sagital da cabeça. Apesar do tratamento de ruído durante a fase de reconstrução das imagens, este ainda está bastante presente nas imagens. Além disso, a inomogeneidade do campo magnético introduz variações locais nas intensidades dos *pixels* em cada corte, dificultando a segmentação do trato vocal.

É relevante mencionar o custo financeiro envolvido para a construção de bases dessa modalidade de imagens; a inexistência de um protocolo comum na área para aquisição de imagens do trato vocal; a indisponibilidade pública dos dados de pesquisas anteriores para avaliação comparativa das metodologias.

4.2 Segmentação Automática dos Contornos do Trato Vocal

Aplicaremos curvas de nível (*level set functions*) para segmentar o trato vocal do restante da imagem. Essa metodologia é conhecida por ser robusta na extração de regiões conexas e na presença de artefatos em uma variedade de aplicações. Esse método de segmentação será conjugado com as similaridades entre imagens consecutivas, propagando os resultados e permitindo uma extração de contornos mais acurada.

Sistematizamos os principais passos da segmentação no fluxograma a seguir:

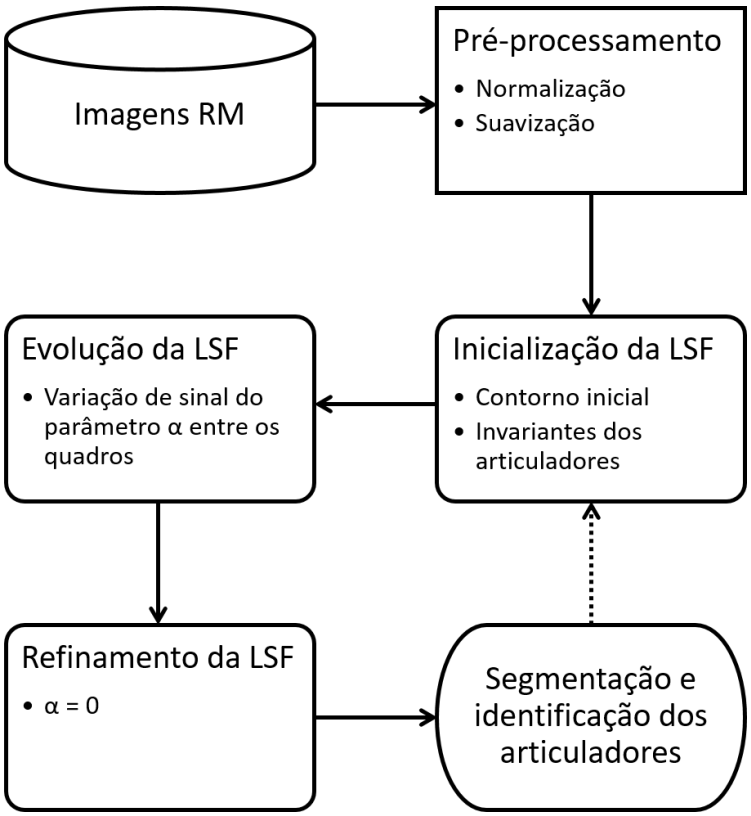


Figura 4.2: *Fluxograma da metodologia de segmentação automática dos contornos do trato vocal.*

4.2.1 Pré-processamento das imagens

Inicialmente é necessário aplicar a transformação de escala das intensidades dos pixels, que leva a representação armazenada em disco à representação para armazenamento em memória, conforme padrão DICOM. Cada imagem possui o metadado *Rescale Intercept (RI)* (0028,1052) e o metadado *Rescale Slope (RS)* (0028,1053), campos estes armazenados no formato DICOM.

A transformação T é aplicada sobre todos os *pixels* da imagem I e é dada por

$$T(x, y) = I(x, y) \cdot RS + RI.$$

As intensidades foram normalizadas linearmente, considerando Min e Max, respectivamente, como a mínima e a máxima intensidades da imagem I, bem como novoMin e novoMax os novos limites mínimo e máximo.

$$I(x, y) = (I(x, y) - \text{Min}) \cdot \frac{\text{novoMax} - \text{novoMin}}{\text{Max} - \text{Min}} + \text{novoMin}$$

Finalmente, aplicamos às imagens o filtro gaussiano G_σ, para suavização de artefatos e ruído. Empiricamente, o σ adequado para essa modalidade de imagens varia de 0,8 a 2,4.

4.2.2 Inicialização das LSF

Três curvas de nível são inicializadas e evoluídas em paralelo para cada imagem, partindo do estado basal do trato vocal. Elas compreendem individualmente a região acima da cavidade aérea, abaixo da cavidade aérea e a parede da faringe. É suficiente que os contornos iniciais estejam próximos das regiões, mesmo não representando perfeitamente as bordas. Alguns pontos desses contornos são escolhidos para identificar os limites das estruturas articulatórias. Tanto os contornos de referência do trato vocal quanto os pontos invariantes das estruturas articulatórias são passados ao modelo uma única vez por falante e perfazem todo o conhecimento *a priori*.

Observamos, ainda, que a evolução do delineamento do trato vocal por meio de três curvas, e não somente uma, foi motivada por duas razões: a dinâmica dos articuladores ser refinada (e.g. a língua é um articulador que exibe movimento que leva a várias possibilidades de constrição); isolar as regiões permitiria, se necessário, adotar estratégias específicas de evolução (e.g. sinal de α diferente).

4.2.3 Segmentação por curvas de nível com Distância Regularizada

Nos métodos convencionais, uma LSF pode evoluir com anomalias causadas por erros numéricos e de estabilidade da curva. Uma técnica usual para evitar tais irregularidades é reiniciar a curva, isto é, suspendendo a sua evolução natural e remodelando a LSF como uma função de distância. No entanto, como pontuado por Gomes e Faugeras (2000), este se mostra um conflito entre teoria e prática, além de introduzir outras dificuldades, como a condição para reiniciar a LSF.

Li *et al.* (2010) ampliaram o conceito de curvas de nível, propondo uma formulação variacional. Os autores introduziram um termo de regularização da distância que torna desnecessário reiniciar a LSF. (Antecipando ao leitor, a evolução da curva de nível se dará como o fluxo do gradiente que minimiza o funcional de energia.)

Seja $\phi : \Omega \to \mathbb{R}$, denotando-se $\phi(\mathbf{x}, t)$ com $\mathbf{x}$ a componente espacial e $t \geq 0$ a componente temporal. Define-se um funcional[2] de energia:

$$\mathcal{E}(\phi) = \mu\mathcal{R}_p(\phi) + \lambda\mathcal{L}_g(\phi) + \alpha\mathcal{A}_g(\phi) \tag{4.1}$$

em que $\mathcal{R}_p(\phi)$ é um termo de regularização da curva de nível com uma função de potencial p que força o módulo do gradiente a um dos pontos de mínimo da curva de nível; $\mu, \lambda > 0$ e $\alpha \in \mathbb{R}$ são constantes; $\mathcal{L}_g(\phi)$ é a integral de linha ao longo da curva de nível zero; $\mathcal{A}_g(\phi)$ corresponde ao peso dado para a área da região de interesse (interior da curva de nível zero) – este termo acelera a evolução da LSF quando o contorno inicial está distante da região de interesse. Tais termos são definidos pelos funcionais:

$$\mathcal{R}_p(\phi) = \int_\Omega p(|\nabla\phi|)d\mathbf{x} \tag{4.2}$$

$$\mathcal{L}_g(\phi) = \int_\Omega g\delta(\phi)|\nabla\phi|d\mathbf{x} \tag{4.3}$$

$$\mathbb{A}_g(\phi) = \int_\Omega gH(-\phi)d\mathbf{x} \tag{4.4}$$

Para a definição das funções g, δ de Dirac e H de Heaviside, vide o Capítulo 2. Observamos, ainda, que $\delta(\phi)$ é zero, exceto quando considerado o nível zero da LSF.

Li *et al.* propõem uma função p de potencial duplo para o termo de regularização

$$p(s) = \begin{cases} \frac{1}{(2\pi)^2}(1 - \cos(2\pi s)) & \text{se } s \leq 1 \\ \frac{1}{2}(s-1)^2 & \text{se } s \geq 1 \end{cases}$$

Essa função p tem dois pontos de mínimo: $s = 0$ e $s = 1$. O objetivo dela é manter a propriedade da distância $|\nabla\phi| = 1$ somente em uma vizinhança da curva de nível zero, para garantir a acurácia da evolução dela. A LSF é constante, com $|\nabla\phi| = 0$, em regiões distantes da curva de nível zero, promovendo a suavização da curva.

Substituindo em eq. (4.1) as eqs. (4.2) a (4.4), temos a seguinte aproximação para o funcional

[2]Em nosso contexto, um *funcional* é um mapeamento $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, em que o espaço $\mathcal{X}$ é um conjunto de funções com certas propriedades (por exemplo, contínuas, suaves etc.).

de energia:

$$\mathcal{E}_\epsilon(\phi) = \mu \int_\Omega p(|\nabla\phi|)d\mathbf{x} + \lambda \int_\Omega g\delta_\epsilon(\phi)|\nabla\phi|d\mathbf{x} + \alpha \int_\Omega gH_\epsilon(-\phi)d\mathbf{x} \tag{4.5}$$

Finalmente, a evolução da curva de nível ϕ será obtida com o fluxo do gradiente que minimiza o funcional de energia $\mathcal{E}_\epsilon(\phi)$.

4.2.4 Fluxo do gradiente para minimização da energia

De acordo com Aubert e Kornprobst (2006), citado por Li *et al.*, uma forma de minimizar um funcional de energia $\mathcal{F}(\phi)$ é encontrar a solução do estado estacionário da equação do fluxo do gradiente:

$$\frac{\partial\phi}{\partial t} = -\frac{\partial\mathcal{F}}{\partial\phi}$$

sendo $\frac{\partial\mathcal{F}}{\partial\phi}$ a derivada de Gâteaux do funcional $\mathcal{F}(\phi)$.
A derivada de Gâteaux de um funcional

$$\mathcal{F}(\phi) = \int_\Omega L\left(\mathbf{x}, \phi(\mathbf{x}), \nabla\phi(\mathbf{x})\right) d\mathbf{x}$$

é definida por

$$\frac{\partial\mathcal{F}}{\partial\phi} = \frac{\partial L}{\partial\phi}(\mathbf{x}, \phi, \nabla\phi) - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\frac{\partial L}{\partial\phi_{x_i}}(\mathbf{x}, \phi, \nabla\phi)\right) \tag{4.6}$$

em que ϕ_{x_i} representa $\frac{\partial\phi}{\partial x_i}$. Aplicando a eq. (4.6) às eqs. (4.2) a (4.4), temos:

$$\begin{aligned}
\frac{\partial\mathcal{R}_p}{\partial\phi} &= \frac{\partial}{\partial x}\left(p' \cdot \frac{\phi_x}{\sqrt{\phi_x^2+\phi_y^2}}\right) + \frac{\partial}{\partial y}\left(p' \cdot \frac{\phi_y}{\sqrt{\phi_x^2+\phi_y^2}}\right) - 0 \\
&= \nabla \cdot \left(p' \cdot \frac{\phi_x}{\sqrt{\phi_x^2+\phi_y^2}}, p' \cdot \frac{\phi_y}{\sqrt{\phi_x^2+\phi_y^2}}\right) \\
&= \operatorname{div}\left(p'\left(|\nabla\phi|\right)\frac{\nabla\phi}{|\nabla\phi|}\right)
\end{aligned} \tag{4.7}$$

$$\begin{aligned}
\frac{\partial\mathcal{L}_g}{\partial\phi} &= \frac{\partial}{\partial x}\left(g\delta(\phi)\frac{\phi_x}{\sqrt{\phi_x^2+\phi_y^2}}\right) + \frac{\partial}{\partial y}\left(g\delta(\phi) \cdot \frac{\phi_y}{\sqrt{\phi_x^2+\phi_y^2}}\right) - 0 \\
&= \delta(\phi) \cdot \operatorname{div}\left(g\frac{\nabla\phi}{|\nabla\phi|}\right)
\end{aligned} \tag{4.8}$$

$$\frac{\partial\mathcal{A}_g}{\partial\phi} = \alpha g\delta(\phi), \text{ haja vista que a função } \delta, \text{ por definição, é a derivada da função } H. \tag{4.9}$$

Substituindo as eqs. (4.7) a (4.9) na equação do fluxo do gradiente originada da eq. (4.5):

$$\frac{\partial\phi}{\partial t} = \mu \cdot \operatorname{div}\left(p'\left(|\nabla\phi|\right)\frac{\nabla\phi}{|\nabla\phi|}\right) + \lambda\delta_\epsilon(\phi) \cdot \operatorname{div}\left(g\frac{\nabla\phi}{|\nabla\phi|}\right) + \alpha g\delta_\epsilon(\phi), \tag{4.10}$$

Discutiremos a implementação da solução numérica da eq. (4.10) a seguir.

4.2.5 Implementação

Li *et al.* implementam o método de segmentação com a regularização da distância utilizando o método de diferenças finitas. Consideraremos uma LSF $\phi(x, y, t)$, com as derivadas espaciais e temporal como definidas no Capítulo 2.

Consideremos um ponto (i, j) da imagem. Esse ponto cruzará o nível zero se $\phi(i-1, j)$ e $\phi(i+1, j)$ têm sinais opostos ou $\phi(i, j-1)$ e $\phi(i, j+1)$ têm sinais opostos. O conjunto de todos os pontos que anulam a LSF será denotado por Z. Construir-se-á uma banda estreita B em função da vizinhança N (3 x 3) centrada no ponto (i, j):

$$B = \bigcup_{(i,j)\in Z} N_{i,j}.$$

Para tanto, o algoritmo compreende as seguintes fases:

1. Iniciar a LSF com uma curva ϕ_0. A banda B será $B^0 = \bigcup_{(i,j)\in Z^0} N_{i,j}$, onde Z^0 é o conjunto dos pontos que anulam a ϕ^0;
2. Atualizar a LSF: $\phi_{i,j}^{k+1} = \phi_{i,j}^{k} + \tau L(\phi_{i,j}^{k})$ na banda B^k, sendo $L(\phi)$ o membro direito da eq. (4.10);
3. Atualizar a banda: determinar o conjunto Z^{k+1} dos pontos que anulam $\phi_{i,j}^{k+1}$ em B^k. Isto é, $B^{k+1} = \bigcup_{(i,j)\in Z^{k+1}} N_{i,j}$;
4. Atualizar os valores dos *pixels* que estão na banda: para todo (i, j) em B^{k+1}, mas não em B^k, $\phi_{i,j}^{k+1} \leftarrow h$ se $\phi_{i,j}^{k} > 0$, caso contrário, $-h$, sendo h uma constante;
5. Detectar o término: se o conjunto dos pontos que anula a LSF não variar após um certo número de iterações, ou um número máximo de iterações for atingido, então o processo deve parar; caso contrário, volta-se à fase 2.

Para fins de implementação, adotou-se:

1. $\Delta x = \Delta y = 1$;
2. $\Delta t = 5$;
3. $\mu = 0.04$;
4. $\lambda = 5$;
5. $\alpha = \pm 1.3$.

A escolha de Δt visa a satisfazer à condição de Courant-Friedrichs-Lewy para um funcional $\mathcal{F}$:

$$\Delta t \leq \frac{\min(\Delta x, \Delta y)}{\max|\mathcal{F}_{ij}|}.$$

Os experimentos realizados por Li *et al.*, assim como os nossos, não evidenciam uma sensibilidade do modelo para os parâmetros μ e λ. Já o parâmetro α é bastante relevante e dependente do tipo de imagem utilizada; o sinal de α é essencial para favorecer a contração ou a expansão da LSF.

Mencionamos, ainda, que a utilização de três curvas de nível para a segmentação do trato vocal permite uma estabilidade numérica superior à evolução de uma única curva que compreendesse o trato vocal integralmente. Devido ao paralelismo computacional, o tempo levado para o processamento não é prejudicado. O tempo total para o processamento da evolução de cada quadro é da ordem de 10 s.[3]

[3] A nossa implementação foi inteiramente realizada em MATLAB 2016a numa máquina de configuração Intel(R) Core(TM) i7-5500U CPU @ 2.40 GHz com 8 GB de memória RAM.

4.2.6 Coerência temporal na segmentação do trato vocal

Uma vez que o trato vocal foi delineado em uma imagem, o processo se repetirá ao longo das subsequentes. Com a posse do contorno do trato vocal, as imagens seguintes se beneficiam dos resultados de segmentações anteriores, aumentando a eficiência e precisão do processo de segmentação.

Após k interações, terminada a segmentação, ϕ_k será ϕ_0 da próxima imagem, isto é, o modelo é reinicializado com o resultado da segmentação anterior.

A importante vantagem no uso dessa modelagem é capturar a variabilidade dos contornos a partir de similaridades entre as imagens, sem a necessidade de amostras para treinamento, devido à coerência temporal. Com a representação de curvas de nível, a informação *a priori* do contorno foi utilizada para condicionar a segmentação apenas na primeira imagem, encapsulando essa informação no termo evolutivo da curva.

4.2.7 Identificação das estruturas articulatórias

A variabilidade de configurações das estruturas articulatórias do trato vocal no processo de fala, especialmente as originadas por oclusões, é capturada no processo de minimização do funcional de energia externa. Este direciona a expansão ou contração da curva de nível para a região de interesse, partindo da localidade das estruturas na segmentação anterior. Cinco pontos invariantes são manualmente marcados somente na imagem que inicializa o método:

- P_1: na arcada dentária superior, separando o lábio superior do palato duro;
- P_2: no palato duro, separando-o do véu palatino;
- P_3: na arcada dentária inferior, separando o lábio inferior da parte anterior da língua;
- P_4: no início da epiglote, separando-a da parte posterior da língua;
- P_5: no término da epiglote, separando-a da parede da faringe.

A identificação automática das estruturas do trato vocal é derivada da evolução da curva de nível associada à coerência temporal. As deformações do trato vocal propagam os limites das estruturas articulatórias, sob pontos invariantes, ao longo das segmentações.

4.3 Validação dos resultados

De acordo com Berry (2007), a segmentação é um dos métodos que requer avaliação para determinar se o resultado obtido está próximo daquilo que se considera "verdade". Em tempo, exploremos a dicotomia entre *gold standard* e *ground truth*: enquanto esta se constituiria como a segmentação verdadeira (exata, perfeita), porém inexistente; aquela representa uma segmentação razoável e em condições adequadas para fins de comparação. Em nosso contexto, a "verdade" está circunscrita ao *gold standard*.

A avaliação do método de segmentação pode ser feita, segundo Berry (2007), a partir de técnicas qualitativas, que se referem a comparações visuais do resultado da segmentação com uma segmentação de referência, e técnicas quantitativas, que se referem, brevemente, a acurácia, precisão e eficiência do método.

Como forma de validar os resultados obtidos da aplicação desta metodologia de segmentação do trato vocal, precisaremos realizar segmentações manuais (construção do *gold standard*) a serem utilizadas nas avaliações qualitativa e quantitativa.

A avaliação qualitativa deverá ser feita por um especialista na área de fonoaudiologia, que julgará se o delineamento do trato vocal, bem como os articuladores destacados, são verossímeis.

A avaliação quantitativa envolverá a mensuração das grandezas que detalhamos a seguir.

4.3.1 Métricas de comparação

Nos trabalhos desenvolvidos por Babalola *et al.* (2008); Kohlberger *et al.* (2012); Morey *et al.* (2009), métricas adequadas para comparar regiões segmentadas são objeto de discussão. Em nosso contexto, consideraremos as métricas: Jaccard, Dice, Tanimoto, Acurácia, as Taxas de Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos e a distância Hausdorff. Tais métricas avaliam a segmentação obtida em relação a uma segmentação manual, em termos de áreas e distâncias entre os contornos. Iremos compará-las para avaliar a segmentação quantitativamente.

Jaccard é definido por:

$$J = \frac{|A_a \cap A_m|}{|A_a \cup A_m|}.$$

Dice é definido por:

$$D = \frac{2\,|A_a \cap A_m|}{|A_a| + |A_m|},$$

em que A_{am} corresponde à intersecção das áreas envolvidas pelos contornos da segmentação automática A_a e da manual A_m, de dimensões $r \times s$. D sempre está entre 0 e 1; quanto mais elevado D, melhor a correspondência entre as segmentações. Idem para J.

Sejam os falsos positivos, os falsos negativos, os verdadeiros positivos e os verdadeiros negativos, respectivamente, definidos por:

$$FP = |A_a| - |A_a \cap A_m|$$
$$FN = |A_m| - |A_a \cap A_m|$$
$$VP = |A_a \cap A_m|$$
$$VN = |\mathbf{1_{rs}}| - |A_a \cup A_m|\,.$$

Em função dessas medidas, definimos as métricas a seguir:

$$\text{Tanimoto} = \frac{VP}{VP + FP + FN}$$

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + FN + VN}.$$

Denotaremos a Taxa de Verdadeiros Positivos por TVP, a Taxa de Verdadeiros Negativos por TVN, a Taxa de Falsos Positivos por TFP e a Taxa de Falsos Negativos por TFN.

$$TVP = \frac{VP}{VP + FN}$$

$$TVN = \frac{VN}{FP + VN}$$

$$TFP = \frac{FP}{FP + VN}$$

$$TFN = \frac{FN}{VP + FN}$$

A taxa de verdadeiros positivos é também conhecida por *sensibilidade*, pois mede a proporção de verdadeiros positivos corretamente avaliados. A taxa de verdadeiros negativos é também conhecida por *especificidade*, pois mede a quantidade de verdadeiros negativos corretamente avaliados.

A distância de Hausdorff H entre os dois conjuntos de pontos A e M pode ser obtida em quatro passos: a) calculam-se todas as menores distâncias de cada ponto de A para todos os pontos em M; b) H_{am} é a maior das distâncias obtidas de A em M; c) repete-se o passo a), calculando-se as menores distâncias de M para A; d) repete-se o passo b) em que H_{ma} é a maior das distâncias

obtidas de M em A. Finalmente,

$$H = \max\{H_{am}, H_{ma}\}.$$

A distância H, pois, nos fornece a máxima distância euclidiana entre os pontos da região segmentada manual e automaticamente.

Capítulo 5

Resultados

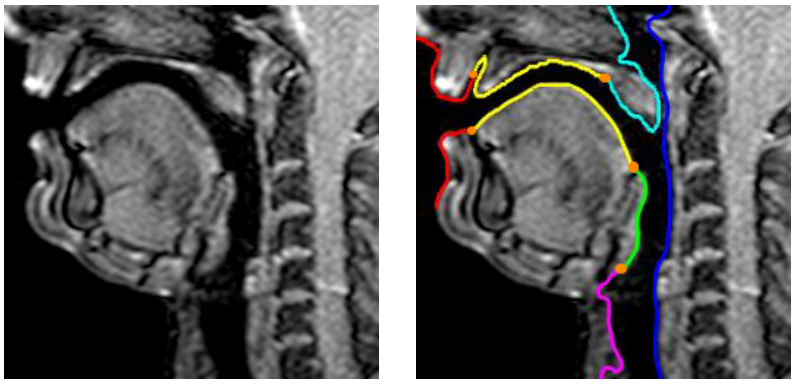
Quando ouvir uma voz dentro de você
dizendo 'Você não pode pintar',
pinte, com toda sua força e de todas
as formas que puder, e a voz será silenciada.

– Vincent van Gogh (pintor, 1853-1890)

Apresentamos os resultados obtidos com a metodologia descrita no capítulo anterior, com vistas à distinção da cavidade aérea dos tecidos do trato vocal, bem como a identificação das estruturas articulatórias.

5.1 Imagens iniciais

Todo o processo envolve partir de imagens de RM do trato vocal, como a imagem 5.1a, e obter segmentações das estruturas articulatórias, como a imagem 5.1b.



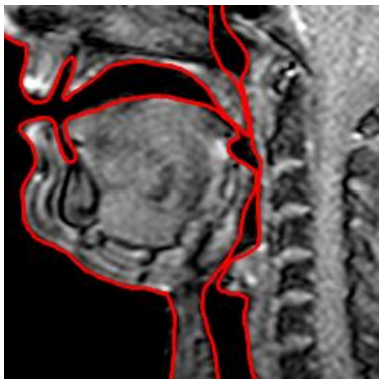
(a) **(b)**

Figura 5.1: *Imagens de RM do trato vocal, adquiridas no Hospital São Paulo (Unifesp), de 0,625 × 0,625 mm^2. A imagem 5.1a corresponde ao nono quadro da série analisada. A imagem 5.1b é resultante da segmentação por curvas de nível. Os pontos invariantes, elencados na seção 4.2.7, são destacados em laranja.*

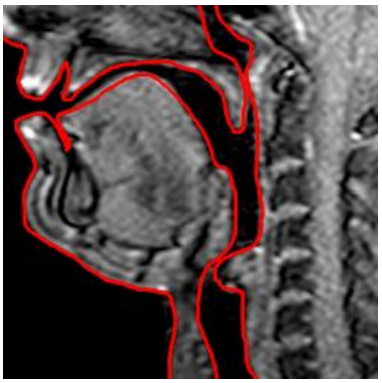
Realizamos a segmentação manual de 10 quadros representativos da articulação do trato vocal, compreendendo estados basais, constritivos e oclusais, os quais apresentamos abaixo. Lembramos que tais segmentações são utilizadas como *gold standard* nas avaliações qualitativa e quantitativa, e foram submetidas à crítica de um especialista da área de fonoaudiologia.

(a) *Quadro 13*

(b) *Quadro 14*

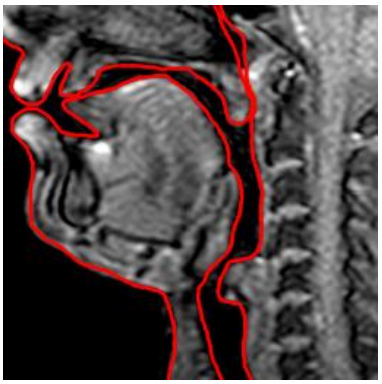


(c) *Quadro 15*

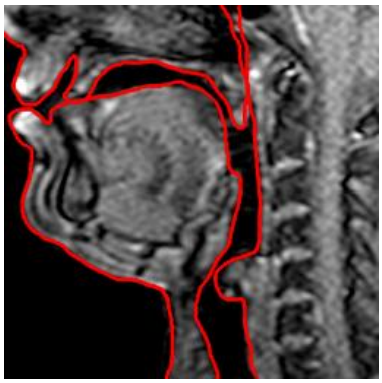


(d) *Quadro 18*

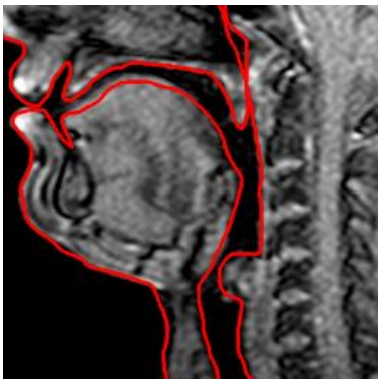
(e) *Quadro 22*



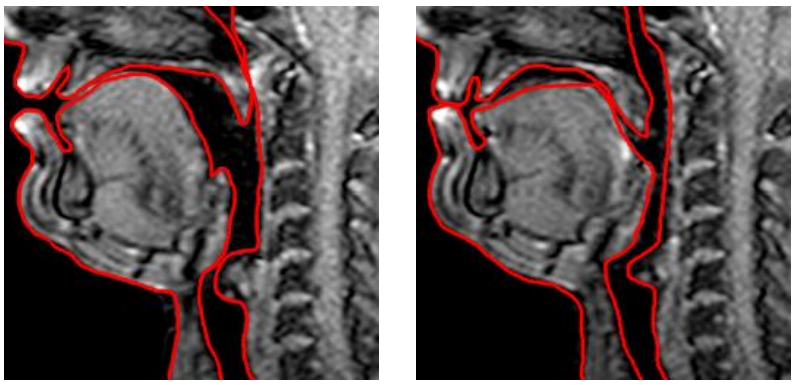
(f) *Quadro 23*



(g) *Quadro 25*



(h) *Quadro 29*



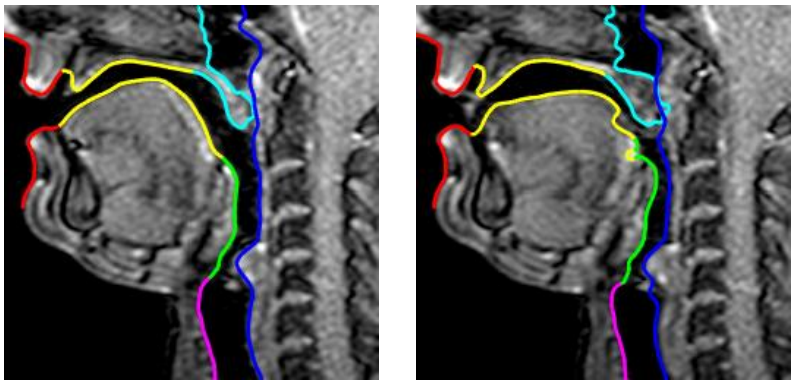
(i) *Quadro 30* **(j)** *Quadro 36*

Figura 5.2: *A precisão da segmentação manual é fundamentalmente dependente da sensibilidade do instrumento de aquisição, bem como da coordenação motora de quem a registra. Ademais, a distinção dos articuladores, principalmente em situações de oclusão, exigem conhecimento anatômico específico para a correta identificação.*

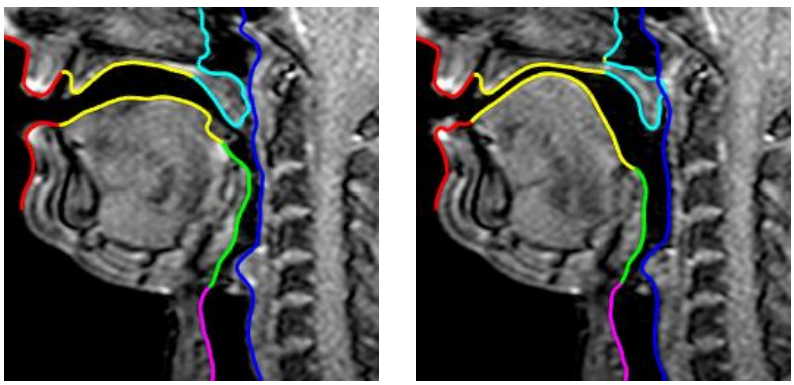
5.2 Avaliação Qualitativa da Evolução da LSF

O modelo é inicializado a partir de uma única segmentação manual e cinco pontos invariantes, sendo esta toda a informação *a priori*.

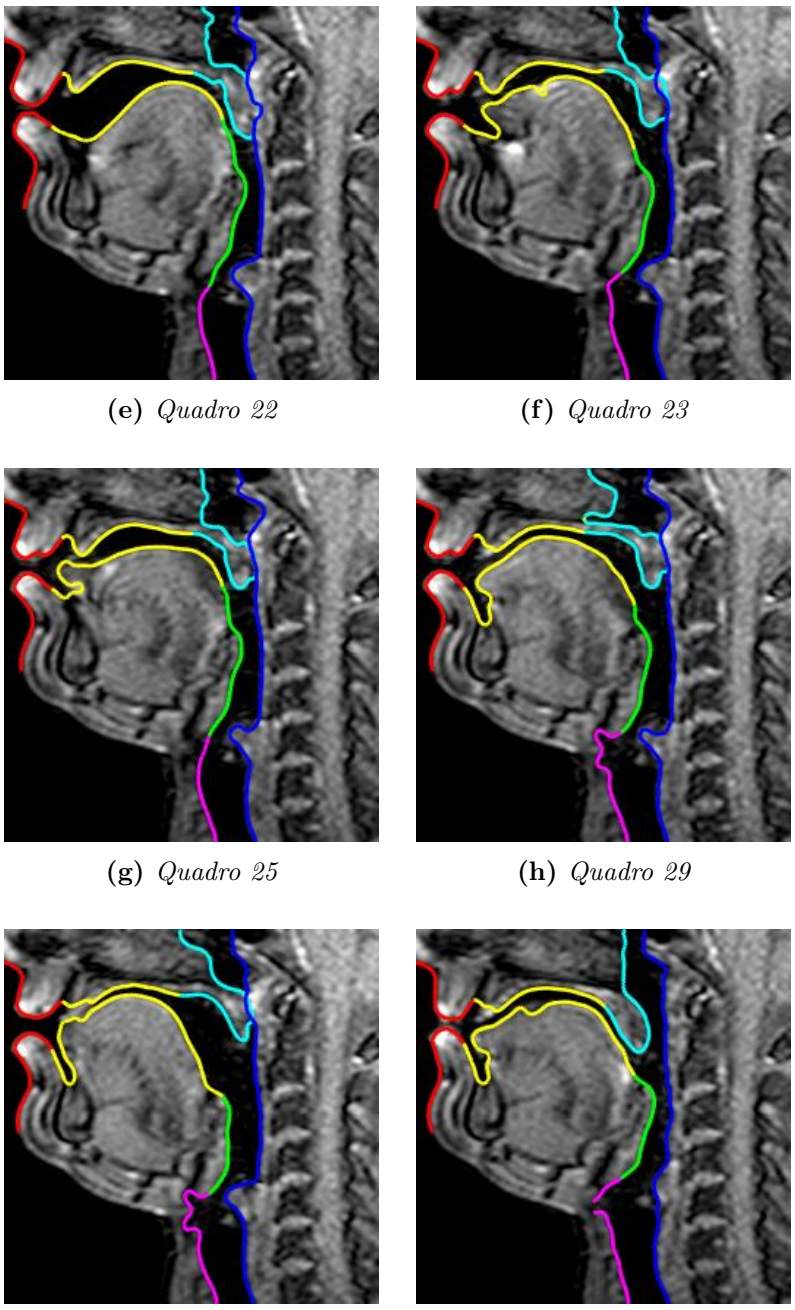
Apresentamos a segmentação do trato vocal ao longo de vários quadros, em que se observam diferentes estados dos articuladores, incluindo situações de oclusões e artefatos gerados por inomogeneidade do campo magnético (especialmente nos lábios).



(a) *Quadro 13* **(b)** *Quadro 14*



(c) *Quadro 15* **(d)** *Quadro 18*



(e) *Quadro 22* **(f)** *Quadro 23*

(g) *Quadro 25* **(h)** *Quadro 29*

(i) *Quadro 30* **(j)** *Quadro 36*

Figura 5.3: *Evolução da LSF ao longo de vários quadros. Observamos, a partir do quadro 22, a presença de artefato no lábio inferior devido à constrição.*

Em vermelho, destacam-se os lábios; em amarelo, o palato duro e a língua; em azul turquesa, o véu palatino; em verde, a epiglote; em azul escuro, a parede da faringe e a glote.

5.3 Avaliação Quantitativa da Evolução da LSF

Considerando-se as métricas apresentadas no Capítulo 4, aplicadas aos mesmos quadros avaliados qualitativamente, obtivemos os seguintes resultados:

Tabela 5.1: *Segmentações na região da faringe*

Quadro	Jaccard	Dice	Tanimoto	Acurácia	TVP	TVN	TFP	TFN	Haussdorf (mm)
13	93%	97%	93%	98%	93%	100%	0%	7%	2,07
14	97%	99%	97%	99%	98%	100%	0%	2%	1,53
15	94%	97%	94%	98%	94%	100%	0%	6%	1,77
18	96%	98%	96%	99%	98%	99%	1%	2%	1,40
22	95%	98%	95%	98%	95%	100%	0%	5%	1,53
23	97%	99%	97%	99%	98%	100%	0%	2%	1,25
25	96%	98%	96%	98%	96%	100%	0%	4%	1,88
29	95%	98%	95%	98%	96%	100%	0%	4%	2,34
30	97%	98%	97%	99%	97%	100%	0%	3%	1,25
36	96%	98%	96%	98%	96%	100%	0%	4%	1,25

Tabela 5.2: *Segmentações na região inferior do trato vocal*

Quadro	Jaccard	Dice	Tanimoto	Acurácia	TVP	TVN	TFP	TFN	Haussdorf (mm)
13	92%	96%	92%	98%	93%	100%	0%	7%	2,58
14	91%	95%	91%	98%	92%	100%	0%	8%	3,06
15	91%	95%	91%	98%	93%	99%	1%	7%	2,65
18	90%	95%	90%	97%	92%	99%	1%	8%	2,17
22	93%	96%	93%	98%	97%	98%	2%	3%	2,17
23	92%	96%	92%	98%	95%	99%	1%	5%	2,25
25	94%	97%	94%	98%	96%	99%	1%	4%	2,17
29	89%	94%	89%	97%	92%	99%	1%	8%	2,34
30	92%	96%	92%	98%	93%	99%	1%	7%	2,17
36	92%	96%	92%	98%	93%	99%	1%	7%	2,58

Tabela 5.3: *Segmentações na região superior do trato vocal*

Quadro	Jaccard	Dice	Tanimoto	Acurácia	TVP	TVN	TFP	TFN	Haussdorf (mm)
13	85%	92%	85%	98%	85%	100%	0%	15%	2,80
14	84%	91%	84%	98%	86%	100%	0%	14%	2,58
15	86%	93%	86%	98%	87%	100%	0%	13%	2,86
18	87%	93%	87%	99%	89%	100%	0%	11%	3,25
22	83%	91%	83%	98%	85%	100%	0%	15%	2,72
23	83%	91%	83%	98%	84%	100%	0%	16%	3,19
25	79%	88%	79%	97%	81%	100%	0%	19%	2,65
29	78%	87%	78%	97%	79%	100%	0%	21%	3,00
30	83%	90%	83%	98%	83%	100%	0%	17%	3,00
36	87%	93%	87%	98%	87%	100%	0%	13%	2,86

5.4 Análise dos resultados

Do ponto de vista qualitativo, o especialista em fonoaudiologia avaliou as segmentações resultantes como ajustadas. Em alguns quadros, nota-se alguma sobreposição dos contornos quando se oprimem o véu palatino e a parede da faringe.[1]

Do ponto de vista quantitativo, em relação às medidas que se utilizam das áreas entre as regiões segmentadas manual e automaticamente – Jaccard, Dice, Tanimoto –, temos uma identificação entre as regiões bastante elevada de maneira geral, mas especialmente para a região da faringe e inferior do trato vocal. A região superior do trato vocal apresenta índices mais baixos, da ordem de 80% em média em contraposição a 90% nas outras regiões. Em termos de sensibilidade e especifidade, o modelo apresentou elevados índices de TVP e TVN em todas as regiões. Mas é importante mencionar que a taxa de falsos negativos aumenta substancialmente para a região superior.

Quanto à correspondência de perímetros entre as segmentações manual e automática, avaliadas pela distância de Haussdorf, temos médias de 1,63 mm para a região da faringe, 2,41 mm para a região inferior e 2,89 mm para a região superior. Sob comparação de valores absolutos, as distâncias são pequenas; em termos relativos, a localidade de algumas estruturas pode estar pontualmente prejudicada – por exemplo, o posicionamento da extremidade da língua.

Por fim, observa-se no uso de curvas de nível a robustez quanto a deformações do trato vocal, que decorre da avaliação em *subpixel* adotada no método. A consistência demonstrada também é relevante, quando comparamos os eventuais erros humanos decorrentes da segmentação manual.

[1] Desta surge uma possibilidade de melhoria: restringir a evolução paralelizada às regiões superior e inferior; o segundo passo seria segmentar a curva referente à parede da faringe, sujeita aos limites impostos pelas anteriores.

Capítulo 6

Conclusões e Perspectivas

Nossas vidas começam a terminar no dia em que permanecemos em silêncio diante de coisas que importam.

– Martin Luther King Jr. (religioso e político, 1929-1968)

As imagens de ressonância magnética em tempo real levaram a progressos sem precedentes no estudo da fala. Na última década, diversas metodologias foram criadas com o objetivo de compreender o todo ou parte do processo articulatório, com diferentes tipos de avaliação, desde a análise somente de vogais até a elaboração de modelos anatômico-geométricos do trato vocal. Em especial, é possível verificar que a imensa maioria dos trabalhos exige uma interação com o usuário em alguma fase, para que este possa fornecer informações que o método não é capaz de identificar (por exemplo, a cavidade aérea em função da constrição da ponta da língua).

No contexto da segmentação dos contornos do trato vocal, observamos o contraste da interface ar-tecido do trato vocal e, assim, também investigamos a aplicação de operadores que destacassem as bordas das estruturas. Estudamos métodos baseados em diferenciais discretos (Roberts, Sobel, Prewitt e o Laplaciano da Gaussiana); alguns operadores morfológicos, especialmente os gradientes morfológicos; o operador de Canny; e a técnica de particionamento baseado em *watershed*. Nenhum deles se mostrou flexível ou robusto o bastante para lidar com oclusões, que são frequentes na dinâmica do trato vocal, e, sobretudo, capturar a consistência intertemporal de formas.

Neste trabalho, desenvolvemos uma metodologia baseada em curvas de nível com distância regularizada, que não requer base de treinamento, tampouco tratamento específico para o idioma do falante, mas conhecimento *a priori* da estrutura do trato vocal do falante na inicialização do método. O texto lido contemplou um número fixo de sentenças, que apresentava variações fonéticas no contexto prosódico e fonológico, subjacentes a características de conversação cotidiana. Os resultados obtidos capturaram o contorno do trato vocal de forma correta, considerando a flexibilidade disponível para lidar com a degeneração, a oclusão e a divisão de estruturas do trato vocal. Nesse sentido, o uso da coerência temporal foi fundamental para a evolução das curvas de nível. Mas cabe observar que, na ausência de informações *a priori*, embora robustas, as curvas de nível não capturam consistentemente os contornos do trato vocal.

Com respeito à identificação das estruturas articulatórias, obtivemos independência relevante da caracterização, por vezes estatística, da variação das formas – o que é considerado na maioria dos outros métodos na literatura. O conhecimento *a priori*, informado somente na imagem inicial e uma única vez por indivíduo, compreende a identificação de cinco pontos invariantes das estruturas articulatórias. Estes imbuem o método a respeito do posicionamento das estruturas, não interferindo na evolução da curva de nível.

Observamos, ainda, a relevância da resolução temporal, de forma a garantir o registro de uma transição gradual de movimento das estruturas articulatórias ao longo dos quadros, e não apenas uma mudança repentina de estado. Ademais, biotipos faciais variados devem ser avaliados antes da prática clínica.

A metodologia resultante deste trabalho poderá ser utilizada em aplicações inovadoras, como

na criação de sistemas para a supressão de sotaque, auxílio à produção da fala em pacientes laringectomizados e terapia de crianças com apraxia da fala.

Referências Bibliográficas

Alvey *et al.* **(2008)** C. Alvey, C. Orphanidou, J. Coleman, A. McIntyre, S. Golding e G. Kochanski. Image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: a comparison using automated image processing. *International Journal of Computer Assisted Radiology and Surgery*, páginas 457 – 464. Citado na pág. 1

Aubert e Kornprobst (2006) Gilles Aubert e Pierre Kornprobst. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations (Applied Mathematical Sciences)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 0387322000. Citado na pág. 17

Avila-García *et al.* **(2004)** M. S. Avila-García, J. N. Carter e R. I. Damper. Extracting tongue shape dynamics from magnetic resonance image sequences. Em *International Conference on Signal Processing*, páginas 288 – 291. Citado na pág. 11

Babalola *et al.* **(2008)** K. Babalola, B. Patenaude, P. Aljabar, J. Schnabel, D. Kennedy, W. Crum, S. Smith, T. Cootes, M. Jenkinson e D. Rueckert. Comparison and evaluation of segmentation techniques for subcortical structures in brain MRI. Em *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*, páginas 409–416. Springer. Citado na pág. 20

Badin *et al.* **(1998)** P. Badin, G. Bailly, M. Raybaundi e C. Segebarth. A three-dimensional linear arti-culatory model based on MRI data. Em *5th International Conference on Spoken Language Processing*, páginas 417 – 420. Citado na pág. 1

Baer *et al.* **(1991)** T. Baer, J. C. Gore, L. C. Gracco e P. W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America*, 90:799 – 828. Citado na pág. 1

Berry (2007) E. Berry. *A Practical Approach to Medical Image Processing*. Series in Medical Physics and Biomedical Engineering. CRC Press. ISBN 9781584888253. Citado na pág. 6, 19

Bresch e Narayanan (2009) E. Bresch e S. Narayanan. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging*, 28(3):323 – 338. Citado na pág. 2, 11

Bresch *et al.* **(2008)** E. Bresch, Y. C. Kim, K. Nayak, D. Byrd e S. Narayanan. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *IEEE Signal Processing Magazine*, 25(3):123 – 132. doi: 10.1109/MSP.2008.918034. Citado na pág. 1

Demolin *et al.* **(1996)** D. Demolin, T. Metens e A. Soquet. Three-dimensional measurement of the vocal tract by MRI. páginas 272 – 275, Philadelphia, USA. Citado na pág. 2

Demolin *et al.* **(2002)** D. Demolin, S. Hassid, T. Metens e A. Soquet. Real-time MRI and articulatory coordination in speech. *Comptes Rendus Biologies*, 325(4):547 – 556. Citado na pág. 1

Dirac (1958) P. A. M. Dirac. *The Principles of Quantum Mechanics*. International Series of Monographs on Physics. Fourth ed. Citado na pág. 7

Eryildirim e Berger (2011) A. Eryildirim e M. Berger. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. Em *European Signal Processing Conference (EUSIPCO)*. Citado na pág. 11

Fontecave e Berthommier (2006) J. Fontecave e F. Berthommier. Semi-automatic extraction of vocal tract movements from cineradiographic data. Em *Interspeech*, páginas 569 – 572. Citado na pág. 1

Gomes e Faugeras (2000) J. Gomes e O. Faugeras. Reconciling distance functions and level sets. *J. Vis. Commun. Image Represent.*, 11(2):209–223. Citado na pág. 16

Gonzalez e Woods (2006) Rafael C. Gonzalez e Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. ISBN 013168728X. Citado na pág. 5

Gregio (2006) F. N. Gregio. Configuração do trato vocal supraglótico na produção das vogais do português brasileiro: dados de imagens de ressonância magnética. Dissertação de Mestrado, PUC-SP, Brasil. Citado na pág. 1

INCA (2015) INCA. Orientações aos pacientes laringectomizados. http://www1.inca.gov.br/conteudo_view.asp?id=111, 2015. Último acesso em 07/12/2016. Citado na pág. 2

Kass *et al.* **(1988)** M. Kass, A. Witkin e D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, páginas 321–331. Citado na pág. 8

Kim *et al.* **(2014)** J. Kim, N. Kumar, S. Lee e S. Narayanan. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data. *International Seminar on Speech Production*. Citado na pág. 2

Kohlberger *et al.* **(2012)** T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann e L. Grady. Evaluating segmentation error without ground truth. Em *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*, páginas 528–536. Springer. Citado na pág. 20

Lammert *et al.* **(2013)** A. Lammert, V. Ramanarayanan, M. Proctor e S. Narayanan. Vocal tract cross-distance estimation from real-time MRI using region of interest analysis. Em *Interspeech*. Citado na pág. 12

Lecuit (1992) V. Lecuit. Sagittal cut to area function transformations: A comparative study. *Mémoire*. Citado na pág. 2

Li *et al.* **(2010)** C. Li, C. Xu, C. Gui e M. Fox. Distance regularized level set evolution and its application to image segmentation. *Image Processing, IEEE Transactions on*, 19(12):3243–3254. Citado na pág. 7, 8, 16

Liu e Udupa (2009) J. Liu e J. Udupa. Oriented active shape models. *IEEE Transactions on Medical Imaging*, 28(4):571 – 584. Citado na pág. 12

Martins (2011) A. L. D. Martins. *Aumento de resolução de imagens de ressonância magnética do trato vocal utilizadas em modelos de síntese articulatória*. Tese de Doutorado, UFSCAR, Brasil. Citado na pág. 1

McRobbie *et al.* **(2003)** Donald W. McRobbie, Elizabeth A. Moore, Martin J. Graves e Martin R. Prince. *MRI from Picture to Proton*. Cambridge University Press. Citado na pág. 5

Morey *et al.* **(2009)** R. Morey, C. Petty, Y. Xu, J. Hayes, H. Wagner, D. Lewis, K. LaBar, M. Styner e G. McCarthy. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage*, 45(3):855–866. Citado na pág. 20

Narayanan *et al.* **(2004)** S. Narayanan, K. Nayak, S. Lee, A. Sethy e D. Byrd. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771 – 1776. Citado na pág. 1

NEMA (2016) Medical Imaging & Technology Alliance NEMA. The DICOM standard. http://dicom.nema.org/standard.html, 2016. Último acesso em 07/12/2016. Citado na pág. 6

Osher e Fedkiw (2003) S. Osher e R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Applied Mathematical Sciences. Springer-Verlag New York. Citado na pág. 8

Osher e Sethian (1988) S. Osher e J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics*, 79: 12–49. Citado na pág. 7

Perkell *et al.* **(1992)** J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta e M. T. T. Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, 92(6):3078 – 3096. Citado na pág. 1

Raeesy *et al.* **(2013)** Z. Raeesy, S. Rueda, J. K. Udupa e J. Coleman. Automatic segmentation of vocal tract images. *IEEE 10th International Symposium on Biomedical Imaging: From Nano to Macro*. Citado na pág. 2, 12

Rueda e Udupa (2011) S. Rueda e J. Udupa. Global-to-local, shape-based, real and virtual landmarks for shape modeling by recursive boundary subdivision. Em *Proceedings SPIE*, volume 7962, páginas 796247–796247–13. Citado na pág. 12

Silva e Teixeira (2015) Samuel Silva e António Teixeira. Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Comput. Speech Lang.*, 33(1):25–46. ISSN 0885-2308. Citado na pág. 12

Stone *et al.* **(2001)** M. Stone, E. P. Davis, A. S. Douglas, M. N. Aiver, R. Gullapalli, W. S. Levine e A. J. Lundberg. Modeling tongue surface contours from cine-MRI images. *Journal of Speech and Hearing Research*, 44(5):1026 – 1040. Citado na pág. 2

Vacavant (2016) Antoine Vacavant. *A Novel Definition of Robustness for Image Processing Algorithms*, páginas 75–87. Springer International Publishing, Cham. ISBN 978-3-319-56414-2. Citado na pág. 6

Vasconcelos *et al.* **(2011)** M. Vasconcelos, S. Ventura, D. Freitas e J. Tavares. Towards the automatic study of the vocal tract from magnetic resonance images. *Journal of Voice*, 25(6):732 – 742. Citado na pág. 12

Whalen *et al.* **(2005)** D. Whalen, K. Iskarous, M. Tiede e D. Ostry. The haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, Hearing Research*, 48:543 – 554. Citado na pág. 1

Ziegler *et al.* **(2012)** W. Ziegler, I. Aichert e A Staiger. Apraxia of speech: Concepts and controversies. *Journal of Speech, Language, and Hearing Research*, 55(5):S1485–S1501. Citado na pág. 2