Universidade de São Paulo
Instituto de Física de São Carlos

Lucimar Sasso Vieira

Morphlet: uma nova família de transformadas *wavelet*
aplicadas ao processo de conversão de voz

São Carlos - SP

2012

Lucimar Sasso Vieira

Morphlet: uma nova família de transformadas
wavelet aplicadas ao processo de conversão
de voz

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Aplicada
Opção Física Computacional.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

Versão Corrigida

São Carlos - SP

2012

AUTORIZO A DIVULGAÇÃO TOTAL OU PARCIAL DESTES TRABALHOS,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE

Ficha catalográfica elaborada pelo Serviço de Biblioteca e Informação do IFSC/USP

Vieira, Lucimar Sasso.

Morphlet: uma nova família de transformadas *wavelet* aplicadas
ao processo de conversão de voz. / Lucimar Sasso Vieira; orientador
Rodrigo Capobianco Guido – São Carlos, 2012.

76 p.

Tese (Doutorado - Programa de Pós-Graduação em Física
Área de Concentração: Física Aplicada - Opção Física Computacional) -
Instituto de Física de São Carlos da Universidade de São Paulo.

1. Processamento de sinais. 2. Conversão de Voz. 3. *wavelets*. 4. *morphlets*.

Título.

Dedico esta tese a minha esposa Driele Francine Nhoato Vieira, a minha filha Valentina Nhoato Vieira e a todos meus familiares e amigos.

Agradecimentos

Agradeço ao meu orientador, Prof Dr Rodrigo Capobianco Guido, ao diretor da Faculdade de Tecnologia de São José do Rio Preto, Prof Dr Waldir Barros Fernandes Junior, aos coordenadores de curso do Centro Universitário do Norte Paulista, Prof Vangrei Fonseca Simão, Profa Katia Cristina de Paulo, Prof Paulo Cesar Fantinato, ao coordenador de curso da União das Faculdades do Grupo FAIMI de Educação, Prof Fernando Renier Gibotti, assim como ao amigo, Prof Sylvio Barbon Junior, pelo suporte neste trabalho. Ao Instituto de Física de São Carlos (IFSC/USP), por fornecer a infra-estrutura necessária para o desenvolvimento deste projeto de pesquisa.

“Na vida é necessário sabedoria para mudar as coisas que podemos mudar; paciência para aceitar as coisas que não podemos mudar e discernimento para saber a diferença entre as duas situações”

Resumo

VIEIRA, L.S. *Morphlet: uma nova família de transformadas wavelet aplicadas ao processo de conversão de voz*. 2011. 76 p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2012.

O presente trabalho de doutorado teve por objetivo a criação de uma nova família de transformadas *wavelet*, chamadas *Morphlets*, que é específica para o processo de conversão de voz. Antes de explicar a criação da *Morphlet*, foi apresentada uma breve revisão bibliográfica sobre o funcionamento da Transformada Discreta Wavelet, sobre os processos de conversão de voz, algoritmos para criação de novas *wavelets*, entre outros tópicos. Em seguida é feita uma descrição detalhada da técnica utilizada para criação das *Morphlets* e, posteriormente, foi apresentado um novo algoritmo para conversão de voz baseado nas *Morphlets*. A criação das *Morphlets*, assim como o algoritmo proposto para conversão de voz baseado nela, inexiste na literatura, até o presente momento. Para testar a eficiência da técnica proposta de conversão de voz usando as *Morphlets* foram realizados testes diversos, principalmente baseados em critérios perceptuais, sendo os resultados obtidos motivadores, o que indicou um avanço na área.

Palavras-chave: Processamento de sinais. Conversão de voz. *Wavelets*. *Morphlets*.

Abstract

VIEIRA, L.S. *Morphlet: a new wavelet transform family applied for voice conversion process*. 2011. 76 p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2012.

The objective of this PhD work is the creation of a new family of wavelet transforms, called Morphlets, particularly designed for voice conversion. Before explaining the creation of this new family of transforms, a brief literature review on the Discrete Wavelet Transform, voice morphing procedures, algorithms to create specific wavelets, among other topics, was presented. This is followed by a detailed description of the technique used to create the new family of wavelets and, then, the algorithm for voice conversion based on this new transform. To date, the creation of Morphlets and the algorithm for voice conversion do not exist in the literature. In order to test the effectiveness of the proposed approach, many tests, mainly based on perceptual criteria, were performed. The results indicate an improvement in the area.

Key-words: Signal processing. Voice conversion. Wavelets. Morphlets.

Lista de Figuras

2.1	Esquema de funcionamento da Transformada <i>Wavelet</i> Discreta ilustrado para um sinal temporal $x_{[1]}$ de n amostras, com decomposição até o terceiro nível.	32
2.2	Relação entre os filtros de análise e síntese. M: operação de mudança aplicada nos filtros. <i>Fonte: (1)</i>	34
2.3	O <i>setup</i> geral. Reproduzido de (4), com permissão do autor.	38
2.4	A mosca, o eletrodo e outros acessórios. Reproduzido de (4), com permissão do autor.	38
2.5	Os 7 padrões de <i>spikes</i> e <i>overlaps</i> . Reproduzido de (4), com permissão do autor.	38
2.6	Interpretação física simplificada do sistema bio-gerador de voz (adaptado de (17), com permissão do autor).	44
4.1	acima: (A): masculino $\rightarrow$ masculino ; abaixo: (B): feminino $\rightarrow$ feminino	63
4.2	acima: (A): masculino $\rightarrow$ feminino ; abaixo: (B): feminino $\rightarrow$ masculino	64
4.3	acima: (A): masculino $\rightarrow$ masculino ; abaixo: (B): feminino $\rightarrow$ feminino	65
4.4	acima: (A): masculino $\rightarrow$ feminino ; abaixo: (B): feminino $\rightarrow$ masculino	66

Lista de Tabelas

2.1	Características das famílias de <i>wavelets</i> mais comuns, incluindo a quantidade de momentos da função <i>wavelet</i> , onde k é o suporte dos filtros. <i>Fonte: (1)</i>	36
2.2	Sub-tipos de fonemas (adaptado de (17), com permissão do autor)	44
4.1	Distância Euclidiana entre a sentença <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mdab0</i> e a <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mjsw0</i>	58
4.2	Distância Euclidiana entre a sentença <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mwbt0</i> e a <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ fjem0</i>	59
4.3	Conversão entre a sentença <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mdab0</i> e a <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mjsw0</i>	61
4.4	Conversão entre a sentença <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ mwbt0</i> e a <i>sal.wav</i> do diretório <i>/timit/ test/ dr1/ fjem0</i>	62

Lista de Abreviaturas

DSP	<i>Digital Signal Processor</i>
DST	<i>Discrete Shapelet Transform</i>
FPGA	<i>Field Programmable Gate Array</i>
MRA	<i>Multi Resolution Analysis</i>
PRFB	<i>Perfect Reconstruction Filter Bank</i>
QMF	<i>Quadrature Mirror Filters</i>
WAV	<i>Waveform Audio Format</i>

Sumário

1	Introdução	25
1.1	Motivação	25
1.2	Objetivos específicos e contribuições do presente trabalho	27
1.3	Organização do Trabalho	27
2	Revisão Bibliográfica	29
2.1	Introdução a Transformada <i>Wavelet</i> Discreta (DWT)	29
2.1.1	Definição da Transformada <i>Wavelet</i> Discreta (DWT)	29
2.1.2	O Princípio de Funcionamento da DWT	31
2.1.3	Como é feita a Localização Tempo-frequência da DWT e da <i>Morphlet</i> .	35
2.1.4	As Principais Famílias de DWTs	36
2.2	Obtenção de novas <i>wavelets</i>	37
2.2.1	A Transformada <i>Spikelet</i>	37
2.2.2	A Transformada <i>Speechlet</i>	42
2.2.3	A Transformada <i>Shapelet</i>	43
2.3	O Sistema bio-gerador de voz e as técnicas usuais para conversão de voz	43
2.3.1	O Sistema bio-gerador de voz	43
2.3.2	Técnicas usuais para conversão de voz	45
3	A Técnica Proposta	49
3.1	Caracterização da <i>Morphlet</i>	49
3.1.1	Os filtros <i>morphlets</i>	50
3.1.2	<i>Major Morphlet</i> e <i>Minor Morphlet</i>	52

3.1.3	<i>A Morphlet</i> de um sinal temporal discreto	53
3.1.4	Cálculo da <i>Morphlet</i> e da <i>Morphlet</i> inversa	54
3.2	O algoritmo de conversão de voz baseado na <i>Morphlet</i>	54
4	Testes e Resultados	57
4.1	Testes de comparação dos resultados com distância Euclidiana	57
4.2	Testes de comparação dos resultados via critério perceptual	60
5	Conclusões e Comentários	67
	Referências	68
	Apêndice I - Publicações durante o Doutorado	73

Capítulo 1

Introdução

Neste capítulo, a motivação, os objetivos específicos e a organização do trabalho são apresentados e explicados com detalhes.

1.1 Motivação

Conversão de voz (1) (2) (3), que também é conhecida como *voice morphing*, *voice transformation*, ou mesmo *voice conversion*, é o processo pelo qual uma sentença pronunciada por determinado locutor, chamado *source speaker*, transforma-se de forma que pareça ser pronunciada por um outro locutor, o *target speaker*. Dentre as aplicações possíveis para essa técnica, destacam-se:

- “enganar” sistemas de autenticação biométrica por voz, verificando a eficiência dos algoritmos autenticador e conversor;
- reproduzir a voz de uma pessoa qualquer em aplicações multimídia e educacionais;
- editar um filme onde determinado ator, ou atriz, já faleceu, sendo que os *dublês* poderiam ter suas vozes convertidas;

- reprodução de um texto escrito, para uso em leitores de *e-mails* por exemplo, através da síntese da voz de um locutor indisponível;
- edição de voz em geral, substituindo locuções indesejadas por outras desejadas;

entre outras aplicações.

Assim como ocorre com a área de conversão de voz, muitos estudos têm sido encontrados na literatura descrevendo técnicas para a criação de novas famílias de *wavelets* (4) (5). Em geral, esses estudos objetivam criar tais *wavelets* para finalidades específicas, tais como reconhecimento de padrões, extração de características, processamento de classes específicas de sinais, entre outros fatores. Nenhum estudo, entretanto, foi apresentado focando na criação de novas *wavelets* para fins de realizar a conversão de voz. Além disso, não há descrição na literatura sobre técnicas de conversão de voz que sejam completamente baseadas em *wavelets*; ao contrário, *wavelets* têm sido, algumas vezes, utilizadas como complemento para técnicas baseadas em sistemas inteligentes (1).

Assim sendo, o estudo apresentado nesta tese consiste de um algoritmo para conversão de voz que está completamente baseado em uma nova família de transformadas *wavelet*, chamadas *Morphlets*. A principal motivação para a realização deste trabalho é, além do desenvolvimento da técnica supra-mencionada, fornecer suporte ao projeto *SpeechAuth*, desenvolvido no laboratório *SpeechLab* do Instituto de Física de São Carlos da Universidade de São Paulo (IFSC/USP), e financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) *.

*<http://www.fapesp.br> - processo nr. 05/00015-1

1.2 Objetivos específicos e contribuições do presente trabalho

Conforme descrito anteriormente, observa-se que os objetivos específicos deste trabalho foram:

- a criação de uma técnica para construção de uma nova família de transformadas *wavelet* que possa ser aproveitada para fins de conversão de voz;
- a utilização da nova família de *wavelets* criada, para obter um novo esquema de conversão de voz que seja prático e eficiente, diferentemente dos métodos atuais;
- avaliar resultados obtidos e compará-los com outras abordagens existentes para conversão de voz, destacando as vantagens do método proposto;
- apresentar o conceito, novo na literatura, de *wavelets* inteligentes, isto é, que podem ser treinadas para servir, em conjunto com o algoritmo proposto, como funções de transferências otimizadas para um fim específico.

1.3 Organização do Trabalho

Em vista dos objetivos previamente detalhados, este trabalho está organizado da forma como segue. No Capítulo 2, é realizada uma revisão na literatura, apresentando tanto a transformada *wavelet* discreta (*Discrete Wavelet Transform* - DWT) e algumas técnicas para criação de novas *wavelets*, quanto alguns métodos utilizados para conversão de voz. No Capítulo 3 é realizada uma descrição detalhada da técnica proposta para criar as *Morphlets* e como elas podem ser utilizadas para fins de conversão de voz. Já o Capítulo 4 apresenta os testes e os resultados

obtidos. Comentários finais e conclusões são descritos no Capítulo 5. Após as referências bibliográficas, o Apêndice I contém as publicações do autor durante o seu curso de doutorado.

Capítulo 2

Revisão Bibliográfica

Neste capítulo os conceitos teóricos necessários ao desenvolvimento do presente projeto são brevemente revistos, englobando a DWT e as técnicas existentes para criação de novas famílias de wavelets, assim como os principais métodos utilizados tradicionalmente para conversão de voz.

2.1 Introdução a Transformada *Wavelet* Discreta (DWT)

2.1.1 Definição da Transformada *Wavelet* Discreta (DWT)

Atualmente existem muitas transformadas que podem ser utilizadas na área de processamento digital de sinais de áudio e voz, sendo, uma delas, a Transformada *Wavelet* Discreta (DWT) (6)(7)(8), a qual visa a conversão de um sinal no domínio do tempo para o domínio tempo-frequência. A decomposição de um sinal $x_{[.]}$ no domínio do tempo, utilizando a DWT, ocorre através da análise de multi-resolução proposta por (9)(10). Dado que $x_{[.]}$ é um vetor em um espaço multidimensional, verifica-se que o mesmo pode ser decomposto em uma soma de outros dois vetores ortogonais, cada um deles pertencendo a um sub-espço vetorial (11). Dessa

forma, o vetor $\vec{x}$ pode ser representado por $\vec{x} = \vec{a} + \vec{d}$, no qual:

$$\vec{a} = \sum_{k=0}^{\frac{n}{2}-1} \langle \vec{x}, \vec{v}_k \rangle \vec{v}_k \quad \text{e} \quad \vec{d} = \sum_{k=0}^{\frac{n}{2}-1} \langle \vec{x}, \vec{w}_k \rangle \vec{w}_k \quad ,$$

ou seja,

- $\vec{a}$ é a projeção de $\vec{x}$ em um sub-espaço V , com uma base de $\frac{n}{2}$ vetores;
- $\vec{d}$ é a projeção de $\vec{x}$ em um sub-espaço W , com uma base de $\frac{n}{2}$ vetores;
- $V \perp W \leftrightarrow \vec{a} \perp \vec{d} \quad \text{e} \quad \vec{v}_i \perp \vec{w}_i \leftrightarrow \langle \vec{v}_i, \vec{w}_i \rangle = 0$.

A decomposição anterior do vetor $\vec{x}$ é denominada decomposição nível 1. Já uma decomposição nível 2 é realizada redecompondo o vetor $\vec{a}$ em dois vetores que pertencem a sub-espaços vectoriais ortogonais e assim sucessivamente de forma recursiva para os níveis 3, 4, ..., j onde $j = \log_2(n)$ e n é o comprimento de $\vec{x}$. Essa generalização pode ser descrita por:

$$\vec{x} = \vec{a}_j + \sum_{i=1}^j \vec{d}_i. \quad (2.1)$$

onde:

- $\vec{a}_j$ é a projeção de $\vec{x}$ num sub-espaço V_j , com uma base contendo $\frac{n}{2^j}$ vetores;
- $\vec{d}_i$ é a projeção de $\vec{x}$ num sub-espaço W_i , com uma base contendo $\frac{n}{2^i}$ vetores;
- $V_j \perp W_j \leftrightarrow \vec{a}_j \perp \vec{d}_j$;

O processo anterior equivale a escrever o sinal temporal $x_{[.]}$ como:

$$x_{[.]} = \sum_{k=0}^{\frac{n}{2^j}-1} H_{j,k}(n) \phi_{j,k}(n) + \sum_{t=1}^j \sum_{k=0}^{\frac{n}{2^t}-1} G_{t,k}(n) \psi_{t,k}(n) \quad (2.2)$$

no qual:

- $\phi(n)$ e $\psi(n)$ formam uma base para escrevê-lo;

- $\phi(n) = \sum_{i=0}^k h_i \phi(2n - i)$, definida recursivamente por dilatações e translações de si mesma, é chamada função *scaling* (12);
- $\psi(n) = \sum_{i=0}^k g_i \phi(2n - i)$, também definida de modo recursivo, é chamada função *wavelet*, sendo, por construção, ortogonal a função *scaling*;
- $H_{j,k}(n) = \langle x, \phi_{j,k}(n) \rangle$ e $G_{t,k}(n) = \langle x, \psi_{t,k}(n) \rangle$;
- os filtros $h_{[\cdot]}$ e $g_{[\cdot]}$ são chamados filtros de análise e um filtro com k coeficientes é dito filtro de **suporte** k .

2.1.2 O Princípio de Funcionamento da DWT

Para a decomposição de um sinal no domínio do tempo utilizando a DWT é necessário utilizar dois filtros digitais (13): um filtro $h_{[\cdot]}$, passa-baixas, e um $g_{[\cdot]}$, passa-altas. A aplicação da DWT em um sinal discreto no tempo $x_{[\cdot]}$ é realizada por meio da convolução discreta entre esse sinal e os filtros $h_{[\cdot]}$ e $g_{[\cdot]}$. Os resultados dessas convoluções são dois sub-sinais. Um deles, $x_{low[\cdot]}$, chamado de detalhe, é o resultado da filtragem passa-baixas e o outro, $x_{high[\cdot]}$, chamado de aproximação, é o resultado da filtragem passa-altas, conforme ilustrado na figura 2.1. Deve-se observar que, em um processo de convolução convencional, o resultado da filtragem de um sinal com n amostras resulta em um sinal com mais de n amostras. Na DWT é aplicado o processo de subamostragem por 2 (*downsampling* por 2), diminuindo assim, pela metade, o comprimento do sinal filtrado. Dessa forma, de acordo com a figura 2.1, verifica-se que o resultado da convolução do sinal $x_{[\cdot]}$ com os filtros $h_{[\cdot]}$ e $g_{[\cdot]}$ produz um sinal com o mesmo comprimento do sinal $x_{[\cdot]}$. O processo todo pode ser resumido da seguinte forma:

$$x_{low[\cdot]} = x_{[\cdot]} * h_{[\cdot]} = \sum_{k=0}^{n-1} h_k x_{2n-k} \quad , \quad (2.3)$$

Figura 2.1- Esquema de funcionamento da Transformada *Wavelet* Discreta ilustrado para um sinal temporal $x_{[-]}$ de n amostras, com decomposição até o terceiro nível.

$$x_{high}[\cdot] = x[\cdot] * g[\cdot] = \sum_{k=0}^{n-1} g_k x_{2n-k} \quad . \quad (2.4)$$

Os filtros $h[\cdot]$ e $g[\cdot]$, chamados de filtros de análise, possuem frequência de corte em meia-banda (-3dB) (14), isto é, $\frac{\pi}{2}$, onde π representa a máxima frequência angular. Isso explica a possibilidade da subamostragem por 2, ou seja, o resultado das filtrações são sinais com a metade da máxima frequência do sinal original. Assim pode-se descartar metade das amostras sem perder informações relevantes no sinal, de acordo com o Teorema de Nyquist. Além disso, um filtro poderá ser considerado filtro *wavelet* apenas se as suas respostas em frequência forem 0 em $\omega = \pi$, no caso do passa-baixas, e 0 em $\omega = 0$, no caso do passa-altas (12).

Em relação aos filtros $h[\cdot]$ e $g[\cdot]$, eles possuem respectivamente duas funções associadas para sua obtenção: uma única função *scaling* (ϕ) e uma única função *wavelet* (ψ), de acordo com sua família (conforme será visto mais adiante). Além dos filtros $h[\cdot]$ e $g[\cdot]$, existem outros filtros associados. São os chamados filtros de síntese, representados por $\bar{h}[\cdot]$ e $\bar{g}[\cdot]$. Esses filtros são utilizados para calcular a transformada inversa, ou seja, obter o sinal original a partir do sinal transformado. Os filtros de síntese seguem às relações das equações 2.5, 2.6 e 2.7, para $k = 0, \dots, n-1$.

$$g_k = (-1)^k h_{n-k-1} \quad , \quad (2.5)$$

$$\bar{h}_k = h_{n-k-1} \quad , \quad (2.6)$$

$$\bar{g}_k = (-1)^{k+1} h_k \quad . \quad (2.7)$$

Caso $h[\cdot]$, $g[\cdot]$, $\bar{h}[\cdot]$, e $\bar{g}[\cdot]$ sejam um banco de filtros de reconstrução perfeita (*perfect reconstruction filter bank* - PRFB), as relações da figura 2.2 são satisfeitas (9)(10). Assim, as condições de *anti-aliasing* e *no-distortion* são também satisfeitas.

Em termos práticos, o algoritmo, proposto por Mallat e que está detalhadamente descrito

Figura 2.2- Relação entre os filtros de análise e síntese. M: operação de mudança aplicada nos filtros.
Fonte: (1).

em (12), é utilizado para a obtenção, ou cálculo computacional, da DWT de um sinal. Apenas os filtros $h_{[-]}$ e $g_{[-]}$ são necessários, não sendo necessário a utilização das funções ϕ e ψ . O cálculo se dá da seguinte maneira: é realizada a multiplicação de duas matrizes para cada nível de transformação. Sendo $A_{[-][]}$ a matriz dos coeficientes dos filtros e $x_{[-]}$ o sinal original, então $y_{[-]} = A_{[-][]}x_{[-]}$ corresponde ao sinal no domínio tempo-frequência, ou domínio *wavelet*, onde a disposição dos coeficientes nas matrizes é a seguinte:

$$A_{[-][]} = \begin{pmatrix} h_0 & h_1 & h_2 & \dots & \dots & \dots & h_{k-1} & 0 & 0 & 0 & 0 & \dots & \dots & 0 & 0 \\ g_0 & g_1 & g_2 & \dots & \dots & \dots & g_{k-1} & 0 & 0 & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & h_0 & h_1 & h_2 & \dots & \dots & h_{k-2} & h_{k-1} & 0 & 0 & \dots & \dots & 0 & 0 \\ 0 & 0 & g_0 & g_1 & g_2 & \dots & \dots & g_{k-2} & g_{k-1} & 0 & 0 & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ h_{k-2} & h_{k-1} & 0 & 0 & \dots & \dots & \dots & 0 & 0 & h_0 & h_1 & \dots & \dots & h_{k-4} & h_{k-3} \\ g_{k-2} & g_{k-1} & 0 & 0 & \dots & \dots & \dots & 0 & 0 & g_0 & g_1 & \dots & \dots & g_{k-4} & g_{k-3} \end{pmatrix}, \quad x_{[-]} = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ x_{n-2} \\ x_{n-1} \end{pmatrix} \quad \text{e} \quad y_{[-]} = \begin{pmatrix} y_0 \\ y_{\frac{n}{2}} \\ y_1 \\ y_2 \\ y_{\frac{n}{2}+1} \\ \dots \\ \dots \\ \dots \\ \dots \\ y_{n-1} \\ y_{\frac{n}{2}-1} \end{pmatrix}.$$

O resultado da DWT, ou seja, o sinal transformado é a matriz $y_{[-]} = \{y_0, y_1, y_2, y_3, \dots, y_{n-2}, y_{n-1}\}$. Verifica-se que o sinal transformado $DWT(x_{[-]}) = y_{[-]}$ possui sempre o mesmo comprimento do sinal original $x_{[-]}$, devido a construção da matriz $A_{[-][]}$. Os processos de *downsampling* e *wrap-around* estão embutidos no procedimento de cálculo. Particularmente, o processo de *wrap-around* ocorre quando os coeficientes dos filtros $h_{[-]}$ e $g_{[-]}$ aparecem nas últimas colunas e depois nas primeiras linhas da matriz $A_{[-][]}$.

O mesmo algoritmo proposto por Mallat, que está detalhadamente descrito em (12), pode ser utilizado para o cálculo da IDWT de um sinal. Nesse caso, apenas os filtros $\bar{h}_{[-]}$ e $\bar{g}_{[-]}$ são necessários, não sendo necessária, novamente, a utilização das funções ϕ e ψ . O cálculo

se dá da seguinte maneira: é realizada a multiplicação de duas matrizes para cada nível de transformação inversa. Porém, ao invés de utilizar-se a matriz $A_{[\cdot][\cdot]}$, é utilizada a sua inversa $A_{[\cdot][\cdot]}^{-1}$ multiplicada pelo sinal no domínio *wavelet* $y_{[\cdot]}$. O sinal resultante dessa operação é o sinal original $x_{[\cdot]}$. Também se faz importante a observação de que a matriz $A_{[\cdot][\cdot]}$ é ortogonal, isto é, $A_{[\cdot][\cdot]}^{-1} = A_{[\cdot][\cdot]}^T$, facilitando a inversão da transformada.

Uma característica muito importante da DWT é a quantidade de momentos nulos (12) da função *wavelet*, relativos ao conjunto de filtros *wavelet*. Essa quantidade é que define a frequência de corte em meia-banda dos filtros $\bar{h}_{[\cdot]}$ e $\bar{g}_{[\cdot]}$. A quantidade de momentos nulos significa que, para um sinal que pode ser representado por um polinômio de grau abaixo de X e uma função *wavelet* que tem X momentos nulos, os coeficientes do detalhamento serão nulos. O cálculo do momento nulo ν é dado pela equação $\nu = \sum_{k=0}^{p-1} \gamma_k^m \psi(\gamma_k)$, na qual p representa a quantidade de pontos da função *wavelet*, m o momento desejado e γ cada ponto onde a função pode possuir valor diferente de 0, isto é, $\gamma = \frac{1}{2}r$, sendo que $r \in \mathbb{Z}^+$.

2.1.3 Como é feita a Localização Tempo-frequência da DWT e da *Morphlet*

Conforme visto na definição da DWT, esta toma como entrada um sinal no domínio do tempo e o leva para o domínio tempo-frequência. Verifica-se que, quanto maior o nível de decomposição, menor será a resolução no tempo. Por outro lado, quanto menor o nível de decomposição maior será a resolução no tempo. Assim, quanto maior a resolução no tempo, menor a resolução na frequência, e vice-versa. Verifica-se que **a DWT e a *Morphlet* possuem essa mesma abordagem de relacionamento das informações de tempo e de frequência**. Se o suporte dos filtros utilizados for maior, a resposta em frequência será mais próxima da ideal e, conseqüentemente, os dados sobre as faixas de frequências de determinada sub-banda serão

mais exatos. De modo análogo, se o suporte dos filtros for menor, a localização temporal será mais precisa. Portanto, dependendo da aplicação, o suporte dos filtros deve ser escolhido de acordo com o que se deseja ressaltar: localização no tempo ou localização na frequência. Para a aplicação específica da *Morphlet* no presente trabalho, acredita-se que um equilíbrio entre as resoluções temporal e frequencial seja ideal.

2.1.4 As Principais Famílias de DWTs

Atualmente, existem várias famílias de DWTs utilizadas em experimentos. As principais são: Haar, Daubechies, Symmlets, Coiflets, Vaidyanathan e Beylkin. As diferenças entre elas são o suporte dos filtros e as respostas em frequência e em fase, o que faz com que as funções ϕ e ψ também sofram os reflexos de tais diferenças. Observa-se que todas essas famílias são constituídas por filtros de resposta ao impulso finita (*finite impulse response* - FIR). Na tabela 2.1 e em (1) constam descrições com tais diferenças. Não faz parte do objetivo deste texto uma descrição mais detalhada, que pode ser obtida facilmente nas referências, principalmente em (1).

Tabela 2.1- Características das famílias de *wavelets* mais comuns, incluindo a quantidade de momentos da função *wavelet*, onde k é o suporte dos filtros. *Fonte: (1).*

Família	Suporte (k)	Momentos
Haar	2	1
Daubechies	par e maior que 4	(k/2)
Coiflets	par e múltiplo de 6	(k/2) - 2
Symmlets	par e múltiplo de 8	(k/2) - 1
Beylkin	18	(k/2) - 2
Vaidyanathan	24	(k/2)

2.2 Obtenção de novas *wavelets*

2.2.1 A Transformada *Spikelet*

Existe uma série de textos e de artigos, tais como (4), (15) e (5), que descrevem técnicas para criar DWTs otimizadas para determinados sinais, isto é *matched wavelets*. Em especial, a *Spikelet* é uma DWT que foi criada para possibilitar a distinção de 7 tipos de padrões encontrados em um sinal proveniente do neurônio *H1* do cérebro de uma mosca *Diptera*, recolhido de maneira extracelular, com o auxílio de um eletrodo, quando ela recebe estímulos visuais do meio externo. Detalhes sobre a *Spikelet* estão sendo descritos tendo em vista que, embora a transformada *Morphlet* pretenda ser específica para a função de conversão de voz, não tendo relação direta com reconhecimento de padrões, a *Morphlet* retém, em sua elaboração, relação com a *Spikelet*.

As figuras de 2.3 até 2.5, extraídas, com permissão do autor, de (4), servem para ilustrar o problema de reconhecimento de padrões do neurônio H1, assim como explicam com detalhes o experimento envolvendo a construção da *Spikelet*. Particularmente, na figura 2.3 pode ser observado o *setup* geral, na figura 2.4 pode ser visto a mosca presa dentro de um bastão plástico localizado em frente do monitor de vídeo que gera um padrão de barras, o eletrodo de captura dos sinais neurais e os demais acessórios e, por fim, na figura 2.5 estão os sete padrões de sinais comumente captados pelo eletrodo mediante a presença de imagens particulares que a mosca observa. Tal experimento foi realizado no laboratório *DipteraLab* do IFSC/USP ¹. Na figura 2.5, os 7 padrões de *spikes* e *overlaps*, são chamados, da esquerda para a direita, de *overlap* acentuado à esquerda, *overlap* à esquerda, *spike*, *overlap* à direita, *overlap* acentuado à direita, *overlap* à esquerda e à direita e *spike* deformado/irregular. Como cada um dos *overlaps* menci-

¹<http://dipteralab.ifsc.usp.br>

Figura 2.3- O *setup* geral. Reproduzido de (4), com permissão do autor.

Figura 2.4- A mosca, o eletrodo e outros acessórios. Reproduzido de (4), com permissão do autor.

onados é constituído de sobreposições de *spikes* em diversas posições, a *Spikelet* foi otimizada para corresponder a um *spike*, chamado *spike* padrão, sendo os *overlaps* identificados por meio da da localização dos *spikes* que os definem.

Construção da *Spikelet*

Depois de definido o *spike* padrão, três conjuntos de etapas foram seguidas: as etapas do grupo A (A1, A2 e A3), as etapas do grupo B (B1, B2 e B3), que ocorrem em paralelo com as etapas

Figura 2.5- Os 7 padrões de *spikes* e *overlaps*. Reproduzido de (4), com permissão do autor.

do grupo A, e, finalmente, as etapas do grupo C (C1, C2 e C3) que são posteriores às etapas dos grupos A e B. O processo está descrito detalhadamente a seguir, em cada uma das suas etapas.

- **Etapa A1:** os três grupos de equações abaixo são reunidos de modo a formar um sistema nas incógnitas $h_0, \dots, h_{k-1}$, o que corresponde à definição da transformada *wavelet* de Daubechies,

$$\sum_{t=0}^{k-1} (-1)^t h_t t^b = 0 \quad , \quad \sum_{t=0}^{k-1} h_t = 2 \quad , \quad \sum_{t=0}^{k-1} h_{t+2l} = 2\delta_{0,l} \quad ,$$

onde $b = 0, 1, \dots, \frac{k}{2} - 1$ e $l \in \mathbb{Z}$. O primeiro grupo é o responsável por fazer com que a função *wavelet* tenha $\frac{k}{2}$ momentos nulos. O segundo grupo, constituído por uma única equação, cuida da normalização dos coeficientes e vem do fato de que para a definição das funções *scaling* de um dado formato, a área correspondente embaixo de cada curva deve ser unitária. O terceiro grupo, cuida da ortogonalidade entre os filtros.

- **Etapa A2:** nesta etapa, retira-se a condição de ortogonalidade da transformada de Daubechies, já que esta condição não terá uso na transformada *Spikelet*. Esse fato é um dos motivos que tornaram difícil a inversão da *Spikelet*, o que não acontece com a *Morphlet*, conforme será visto adiante no capítulo 3.

$$\sum_{t=0}^{k-1} (-1)^t h_t t^b = 0 \quad , \quad \sum_{t=0}^{k-1} h_t = 2 \quad .$$

- **Etapa A3:** as equações da etapa A2 são especificadas para um suporte $k = 4$. O sistema resultante desta etapa é linear e possui $\frac{k}{2} + 1 = 3$ equações com $k = 4$ incógnitas:

$$\begin{cases} -1h_3 + 1h_2 - 1h_1 + 1h_0 = 0 \\ -3h_3 + 2h_2 - 1h_1 + 0h_0 = 0 \\ 1h_3 + 1h_2 + 1h_1 + 1h_0 = 2 \end{cases} .$$

- **Etapa B1:** realiza-se a convolução seguida de sub-amostragem por 2, do *spike* padrão com um filtro de suporte 4 com coeficientes genéricos $h_0, \dots, h_3$. O sinal resultante é a projeção do *spike* padrão, designado doravante de vetor $\vec{u}$, num subespaço vetorial W .

- **Etapa B2:** a energia do sinal resultante da etapa B1 anterior, isto é, a projeção de $\vec{u}$ em W , é maximizada, objetivando transferir o máximo de energia do sinal original para o sinal resultante. Para isso, são calculadas as suas derivadas parciais em relação a cada um dos coeficientes

$$\frac{\partial \overrightarrow{E(P_w U)}}{\partial h_0} = 0 \quad , \quad \frac{\partial \overrightarrow{E(P_w U)}}{\partial h_1} = 0 \quad , \quad \frac{\partial \overrightarrow{E(P_w U)}}{\partial h_2} = 0 \quad , \quad \frac{\partial \overrightarrow{E(P_w U)}}{\partial h_3} = 0 \quad .$$

- **Etapa B3:** cada uma das 4 derivadas parciais obtidas na etapa B2 é igualada a 0, formando um sistema linear homogêneo de 4 equações com 4 incógnitas:

$$\begin{cases} -3.4984h_3 + 3.3876h_2 - 3.1079h_1 + 2.6457h_0 = 0 \\ -3.3876h_3 + 3.4876h_2 - 3.3967h_1 + 3.0924h_0 = 0 \\ -3.1079h_3 + 3.3967h_2 - 3.5100h_1 + 3.4000h_0 = 0 \\ -2.6457h_3 + 3.0924h_2 - 3.4000h_1 + 3.5006h_0 = 0 \end{cases} .$$

- **Etapa C1:** reúnem-se as equações obtidas nas etapas A3 e B3. O novo sistema é linear e possui $(\frac{k}{2} + 1) + k = 7$ equações e $k = 4$ incógnitas:

$$\begin{pmatrix} -3.4984 & 3.3876 & -3.1079 & 2.6457 \\ -3.3876 & 3.4876 & -3.3967 & 3.0924 \\ -3.1079 & 3.3967 & -3.5100 & 3.4000 \\ -2.6457 & 3.0924 & -3.4000 & 3.5006 \\ -1 & 1 & -1 & 1 \\ -3 & 2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \end{pmatrix} .$$

- **Etapa C2:** devido à inexistência de solução exata para o sistema da etapa C1, optou-se pela solução através do Método dos Mínimos Quadrados. O sistema é então expresso na forma matricial e cada lado da igualdade é multiplicado pela matriz transposta dos pesos das incógnitas:

$$\begin{pmatrix} -3.4984 & -3.3876 & -3.1079 & -2.6457 & -1 & -3 & 1 \\ 3.3876 & 3.4876 & 3.3967 & 3.0924 & 1 & 2 & 1 \\ -3.1079 & -3.3967 & -3.5100 & -3.4000 & -1 & -1 & 1 \\ 2.6457 & 3.0924 & 3.4000 & 3.5006 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -3.4984 & 3.3876 & -3.1079 & 2.6457 \\ -3.3876 & 3.4876 & -3.3967 & 3.0924 \\ -3.1079 & 3.3967 & -3.5100 & 3.4000 \\ -2.6457 & 3.0924 & -3.4000 & 3.5006 \\ -1 & 1 & -1 & 1 \\ -3 & 2 & -1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} h_0 \\ h_1 \\ h_2 \\ h_3 \end{pmatrix}$$

$$= \begin{pmatrix} -3.4984 & -3.3876 & -3.1079 & -2.6457 & -1 & -3 & 1 \\ 3.3876 & 3.4876 & 3.3967 & 3.0924 & 1 & 2 & 1 \\ -3.1079 & -3.3967 & -3.5100 & -3.4000 & -1 & -1 & 1 \\ 2.6457 & 3.0924 & 3.4000 & 3.5006 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 2 \end{pmatrix}.$$

- **Etapa C3:** a solução obtida implica os coeficientes $h_0, \dots, h_3$ encontrados, que representam o filtro passa-baixas da *Spikelet*. Como o sistema é definido do tipo espelho, ou seja, *Quadrature Mirror Filter* - QMF) (14), os coeficientes do filtro passa-alta respectivo obedecem à equação 2.5, ficando, portanto, a *Spikelet* assim definida:

$$\begin{cases} h_0 = 0.26964482896235847376 & , & g_0 = 0.23524044702452745481 \\ h_1 = 0.76237548312490721614 & , & g_1 = -0.73270322306815560687 \\ h_2 = 0.73270322306815560687 & , & g_2 = 0.76237548312490721614 \\ h_3 = 0.23524044702452745481 & , & g_3 = -0.26964482896235847376 \end{cases}.$$

Reconhecimento dos Padrões com a *Spikelet*

O algoritmo para reconhecimento de padrões baseado na *Spikelet* consiste apenas em analisar a transformada nível 3 do sinal de 32 amostras (a mesma quantidade do *spike* padrão) que se queira classificar. Como o sinal sob análise é amostrado a uma taxa de 44100 Hz, a máxima frequência presente é de 22050 Hz de acordo com o Teorema de Nyquist onde diz que o número de amostras por segundo de um sinal que está sendo discretizado deve ser, no mínimo, o dobro da máxima frequência presente no sinal. Portanto, a decomposição implica:

- amostras 0 a 3 da *Spikelet*: descrevem a faixa de frequências de 0 a 2756,25 Hz com incerteza de 8 pontos em relação ao sinal sob análise pois cada uma dessas quatro amostras equivale a localização temporal de 8 amostras no sinal original;
- amostras 4 a 7 da *Spikelet*: descrevem a faixa de frequências de 2756,25 a 5512,50 Hz com incerteza de 8 pontos em relação ao sinal sob análise pois cada uma dessas quatro amostras equivale a localização temporal de 8 amostras no sinal original;

As duas sub-bandas acima, representadas pelas amostras de 0 a 3 e de 4 a 7, são suficientes para a análise, já que a decomposição de um *spike/overlap* nestas frequências mostra todas as

bordas de subida existentes (região onde a *Spikelet* apresenta maior amplitude). Daí, de acordo com (16), para uma *wavelet* não simétrica, como é o caso da *Spikelet*, as bordas de subida no sinal original correspondem aos pontos onde a transformada assume valor 0 (*zero-crossings*) e, consequentemente, às inversões do sinal entre os seus pontos vizinhos. Especificamente, as inversões de sinal entre as amostras vizinhas da *Spikelet* de um provável *spike/overlap* são:

- *overlap* acentuado à esquerda: inverte entre 1 e 2, 4 e 5;
- *overlap* à esquerda: inverte entre 1 e 2, 4 e 5, 5 e 6;
- *spike*: inverte entre 1 e 2, 4 e 5, 5 e 6, 6 e 7;
- *overlap* à direita: inverte entre 1 e 2, 5 e 6, 6 e 7;
- *overlap* acentuado à direita: inverte entre 1 e 2, 6 e 7;
- *overlap* à esquerda e à direita: inverte entre 1 e 2, 4 e 5, 6 e 7;
- *spike* deformado/irregular: qualquer outra combinação.

2.2.2 A Transformada *Speechlet*

A transformada *Speechlet* (5) consiste numa tentativa de encontrar padrões de sinais de voz, utilizando exatamente a mesma técnica e o mesmo algoritmo da *Spikelet*, trocando somente o sinal de interesse, isto é, o *spike* padrão, pelo sinal de formato adequado para reconhecimento do padrão correspondente. Para fins de reconhecimento e distinção de alguns sons vozeados, isto é, *voiced speech*, tais como as vogais /a/, /e/ e /i/, os resultados foram relativamente satisfatórios, entretanto, devido ao uso do método dos mínimos quadrados para obtenção dos coeficientes dos filtros, não foi possível estabelecer garantias de que a classificação seria sempre eficiente.

2.2.3 A Transformada *Shapelet*

A Transformada Discreta *Shapelet* (15) (*Discrete Shapelet Transform* - DST), criada em 2008, constitui um avanço na *Spikelet* e na *Speechlet*. Particularmente, a DST foi criada para possibilitar uma análise conjunta tempo-frequência-formato de sinais. Isso significa que, além da análise tempo-frequência tradicional, essa família de *wavelets* foi criada para possibilitar uma análise do formato do sinal em questão. A idéia utilizada para construir as *Morphlets* deriva da metodologia utilizada para construir a DST.

2.3 O Sistema bio-gerador de voz e as técnicas usuais para conversão de voz

2.3.1 O Sistema bio-gerador de voz

Segundo (17), o mecanismo de produção de voz, por mais complexo que seja, pode ser descrito de forma pormenorizada como sendo a propulsão de ar pelos pulmões seguida de um processo de equalização, garantido pelo trato vocal e elementos associados, conforme pode ser observado na Figura 2.6. Um fato importante corresponde ao comportamento das pregas vocais, que controlam o fluxo de ar fornecido pelos pulmões, tornando o sinal de excitação quase-periódico, vibrando em uma determinada frequência, ou aperiódico, similar a um sinal ruidoso. No primeiro caso, o período do mesmo será chamado de período de *pitch* e a voz produzida será classificada como *voiced speech*, do contrário a mesma será classificada como *unvoiced speech*, assemelhando-se a um ruído. Os principais exemplos de *voiced speech* são as vogais, já a maioria dos demais sons emitidos pelo ser humano podem ser classificados como

Tabela 2.2- Sub-tipos de fonemas (adaptado de (17), com permissão do autor)

tipo	descrição	exemplo
<i>plosives</i>	<i>unvoiced speech</i> impulsivo	<i>t</i> na palavra <i>top</i>
<i>fricatives</i>	<i>unvoiced speech</i> - presença de uma turbulência - ar entre a língua e os dentes superiores	<i>th</i> na palavra <i>thin</i>
<i>whispers</i>	<i>unvoiced speech</i> com uma barreira nas pregas vocais fazendo com que elas permaneçam parcialmente fechadas	<i>h</i> na palavra <i>he</i>
<i>voiced fricatives</i>	<i>voiced speech</i> misturados com ruído criado na constrição do trato vocal contra o palato e atrás dos dentes	<i>z</i> na palavra <i>zebra</i>
<i>voiced plosives</i>	fonemas vozeados em conjunto com ruído impulsivo criado no trato vocal	<i>g</i> na palavra <i>glass</i>
<i>unvoiced fricatives</i>	idem anterior, porém as pregas vocais não vibram conjuntamente com a fricção	<i>h</i> na palavra <i>human</i>
<i>unvoiced plosives</i>	idem anterior, mas as pregas vocais não vibram conjuntamente com o impulso	<i>b</i> na palavra <i>boat</i>

unvoiced speech. Uma característica importante é obtida com o véu palatino, pois o mesmo controla a passagem do fluxo de ar pelos tratos nasal e vocal, sendo que eles podem refinar a qualidade dos sinais para uma classificação conforme a tabela 2.2.

Figura 2.6- Interpretação física simplificada do sistema bio-gerador de voz (adaptado de (17), com permissão do autor).

2.3.2 Técnicas usuais para conversão de voz

De acordo com (1), os sistemas mais elementares de conversão de voz eram baseados em dois estágios, onde o primeiro utilizava alguma técnica de reconhecimento de voz (*speech-to-text*) para então sintetizar a *morphed speech* com base em alguns parâmetros que modelavam o *target speaker*. Este tipo de sistema é bastante complexo e está sujeito a varios problemas, principalmente porque o estágio inicial de reconhecimento de fala nunca é perfeito. Além disso, a complexidade computacional de um sistema como este é bastante grande, sem contar os artefatos introduzidos.

Mais à frente, outras tentativas foram experimentadas. Arslan e Talkin (18) propuseram um sistema no qual as vozes de ambos os locutores, *source* e *target*, eram segmentadas em fonemas e transcritos por intermédio de um dicionário pré-existente, sendo que a qualidade da voz produzida ficava limitada ao tamanho do dicionário.

Stylianou *et al* (19) sugeriram melhorias neste sistema através do uso de modelos de mistura Gaussianos (20) para modelar os parâmetros espectrais dos locutores. Particularmente, as vozes do *source* e do *target* são alinhadas no tempo, seguido da aplicação do modelo Gaussiano, previamente treinado por intermédio do algoritmo de *expectation-maximization* (EM), para realizar o mapeamento. Esta técnica produz resultados um pouco mais próximos de uma voz natural.

Kain (21) apresentou um algoritmo de mapeamento do envelope espectral da mesma maneira como fizeram Arslan e Talkin, incluindo a predição do erro na tentativa de melhora dos resultados. Orphanidou *et al* (22) usaram *Generative Topographic Mapping*, que é um modelo de mistura Gaussiana não-linear para transformar os parâmetros espectrais do *source* naqueles

do *target*, de acordo com o modelo do trato vocal do locutor.

Um dos problemas mais significantes nas abordagens anteriores é a perda de informação durante o processo de síntese da voz do *target speaker*, especialmente nas frequências mais altas, o que leva a resultados não muito satisfatórios em termos de qualidade dos sons não vocálicos convertidos. Uma abordagem tempo-frequência baseada em *wavelets* foi proposta em (23) e apresentou resultados promissores, já que as características de reconstrução perfeita do sinal sob análise, assim como a baixa perda de informação estavam asseguradas. Mais especificamente, a representação dos sinais de voz em multi-escala faz com que o problema de conversão seja sub-dividido em vários problemas de menores dimensões. Em seguida à aplicação da transformada *wavelet*, uma rede neural tipo *multi-layer perceptron* foi usada para fazer o mapeamento do *source speaker* para o *target speaker*, sub-banda por sub-banda. O algoritmo completo é o que segue.

-
- INÍCIO
 - PASSO 1: escolher os sinais do *source speaker* e do *target speaker*, sendo que ambos devem estar pronunciando a mesma sentença;
 - PASSO 2: alinhar os sinais no tempo, isto é, reamostrá-los de forma que eles contenham o mesmo comprimento;
 - PASSO 3: normalizar os sinais para que eles possuam média 0 e desvio-padrão 1;
 - PASSO 4: aplicar a DWT nível 5, obtendo 6 conjuntos de coeficientes: a aproximação do nível 6, e os detalhes dos níveis 1 até 5;
 - PASSO 5: colocar os detalhes dos níveis 1 e 2 em zero;
 - PASSO 6: normalizar os detalhes dos níveis 3 até 6;

- PASSO 7: para a aproximação do nível 6, e os detalhes dos níveis 3 até 6, treinar uma rede *multilayer perceptron*, de forma que o sinal *source* se transforme, mediante passagem pela rede, no sinal *target*, amostra por amostra;
 - PASSO 8: aplicar a DWT inversa de forma a inverter o passo 4;
 - PASSO 9: desnormalizar o sinal *target* obtido de forma a inverter o passo 3;
 - PASSO 10: testar o sinal de saída obtido.
 - FIM.
-

Embora esse algoritmo seja considerado como sendo um dentre os que fazem parte do estado-da-arte em termos de conversão de voz, os autores não apresentam um estudo sobre a escolha dos filtros *wavelet* utilizados, ao contrário, eles simplesmente adotam o uso de uma família de filtros com determinado comprimento.

Uma outra característica que tem sido percebida (2) (3) é que, normalmente, os algoritmos atuais estão, independentemente da técnica específica na qual estão baseados, concentrados em segmentar os trechos de vozes do *source speaker* e do *target speaker* de forma a separar as partes vozeadas das não vozeadas (24), tratando apenas de modificar as primeiras, já que as conversões das segundas parecem não causar muita diferença na prática. Tal procedimento implica sempre em uma dificuldade adicional, devido, também à necessidade de alinhamento dos fonemas correspondentes do *source* e do *target speaker*.

A técnica proposta neste trabalho, que está detalhadamente descrita no próximo capítulo, baseia-se em um conceito novo, que não requer a separação de trechos vozeados e não vozeados, além de não necessitar de alinhamento entre fonemas do *source* e do *target speaker*, já que as

sentenças são convertidas por inteiro.

Capítulo 3

A Técnica Proposta

Este capítulo apresenta a técnica para construção das Morphlets, como sendo um conjunto de modificações na formulação da DST, além do algoritmo para uso das Morphlets em conversão de voz, entre outros detalhes.

3.1 Caracterização da *Morphlet*

Os elementos que formam a estrutura da *Morphlet*, que é similar à estrutura da DST, são os seguintes:

- um filtro passa-baixas meia-banda com resposta ao impulso finita (FIR) (14), $p_{[\cdot]}$, que apresenta resposta em frequência $P[\omega]$ de forma que $P[\omega = \pi] = 0$, e resposta em fase não necessariamente linear;
- o filtro FIR passa-altas com resposta em frequência espelhada, $q_{[\cdot]}$, definido como $q_k = (-1)^k p_{N-k-1}$;
- o filtro FIR $\bar{p}_{[\cdot]}$, definido como $\bar{p}_k = p_{N-k-1}$;
- o filtro FIR $\bar{q}_{[\cdot]}$, definido como $\bar{q}_k = (-1)^{k+1} p_k$;

- o suporte, ou número de coeficientes, dos filtros, N , ($N \geq 4$), que deve ser par;
- as funções *minor morphlet* e *major morphlet*, caracterizadas a seguir.

Os filtros $p_{[\cdot]}$ e $q_{[\cdot]}$ formam o par de filtros de análise e os filtros $\bar{p}_{[\cdot]}$ e $\bar{q}_{[\cdot]}$ formam o par de filtros de síntese. O primeiro é utilizado para transformar o sinal original do domínio do tempo para o domínio tempo-frequência, enquanto que o segundo par é utilizado para inverter a transformação, recuperando o sinal original a partir do transformado. As condições impostas acima para $p_{[\cdot]}$, $q_{[\cdot]}$, $\bar{p}_{[\cdot]}$, $\bar{q}_{[\cdot]}$ e N forçam os filtros a formar um PRFB, isto é, as condições de *anti-aliasing* e *no-distortion* são satisfeitas.

3.1.1 Os filtros *morphlets*

Uma vez que os coeficientes de um dos filtros FIR especificados acima tenham sido determinados, a *Morphlet* está completamente definida. Para determiná-los deve ser seguido o algoritmo seguinte, com cinco sequências de passos, A, B, C, D, E e F. Uma descrição detalhada de cada passo é apresentada a seguir, visando determinar o filtro $p_{[\cdot]}$.

- **Passo A:** O filtro $p_{[\cdot]}$ deve ser especificado de forma a ter energia unitária, visando conservar, no sinal transformado, a energia do sinal original:

$$\sum_{k=0}^{N-1} p_k^2 = 1 \quad . \quad (3.1)$$

- **Passo B:** As condições de ortogonalidade são impostas entre as translações dos filtros:

$$\sum_{k=0}^{N-1} p_k p_{k+2l} = \delta_{0,l} \quad , \quad (3.2)$$

δ sendo o delta Dirac e $l \in \mathbb{Z}$.

- **Passo C:** O sistema deve possuir, no mínimo 1 (condição necessária para caracterização

de uma *wavelet*), e no máximo $\frac{N}{2} - 1$ (máximo nesse caso, visando não obter um sistema com mais de N equações), momentos nulos:

$$\sum_{k=0}^{N-1} (-1)^k p_k k^b = 0 \quad , \quad (3.3)$$

para $b = 0$ ou 1 ou ... ou $\frac{N}{2} - 2$, conforme o número de momentos escolhido.

- **Passo D:** A *Morphlet* nível máximo de $s_{i[\cdot]}$ e de $t_{i[\cdot]}$, ($Morphlet_{max}(s_{i[\cdot]})$) e ($Morphlet_{max}(t_{i[\cdot]})$), sendo $s_{i[\cdot]}$ e $t_{i[\cdot]}$ respectivamente o i -ésimo *source signal* e o i -ésimo *target signal* correspondente, devem ser calculadas, conforme explicado anteriormente durante a revisão bibliográfica sobre *wavelets*, usando os filtros $p_{[\cdot]}$ e $q_{[\cdot]}$, de suporte N , com coeficientes genéricos $p_0, \dots, p_{N-1}$ e $q_0, \dots, q_{N-1}$.

Observa-se que pode haver de 1 até $\frac{N}{2} - 1$ *source* e *target signals* correspondentes, dependendo, respectivamente, do número de momentos escolhido no passo C, que pode variar de 1 até $\frac{N}{2} - 1$. Quanto maior o número de momentos, mais a curva de resposta em frequências do filtro *Morphlet* se aproxima da ideal, por outro lado, quanto menor o número de momentos, maior a quantidade de *source* e *target signals*, isto é, maior é o “treinamento” que a *Morphlet* recebe para “aprender” a função de transferência entre os locutores.

Por fim, deve-se impor a condição

$$Morphlet_{max}(s_{i[\cdot]}) \cdot Morphlet_{max}(t_{i[\cdot]}) = 0 \quad , \quad (3.4)$$

isto é, o produto escalar das duas *Morphlets* é nulo, o que equivale a impor uma

condição de ortogonalidade entre ambas. A relação

$$p_k = -1^{k+1} q_{N-k-1}$$

deve ser utilizada de forma que ambas as *Morphlets* sejam funções de $p_{[\cdot]}$ somente.

- **Passo E:** As equações dos passos A, B, C e D devem ser agrupadas, resultando em um sistema não linear de N equações e N incógnitas.
- **Passo F:** O sistema deve ser resolvido utilizando qualquer método numérico iterativo propício, tal como o método de Newton (25), obtendo-se assim o conjunto de coeficientes p_i , isto é, o filtro $p_{[\cdot]}$.

3.1.2 *Major Morphlet e Minor Morphlet*

Tendo em vista as condições de resposta em frequência e suporte finito que foram impostas aos filtros acima, esses originam funções contínuas de comprimento finito, de formato particular, isto é, *morphlets*, da mesma forma como ocorre com qualquer DWT. Especificamente, tem-se duas funções que são chamadas de *major morphlet*,

$$\Gamma(x) = \sum_k p_k \Gamma(2x - k) \quad ,$$

e *minor morphlet*,

$$\Theta(x) = \sum_k q_k \Gamma(2x - k) \quad ,$$

como sendo as funções *scaling* e *wavelet* (12) de qualquer DWT, respectivamente. Nem Γ nem Θ são de fato necessárias para calcular a *Morphlet* de um sinal, entretanto, elas têm um papel importante em termos de decomposição dos sinais sob análise, já que a *Morphlet* nível j de um sinal $f(x)$, consiste em escrevê-lo como combinação linear destas duas funções, de acordo com

a equação 3.5.

$$f(x) = \underbrace{\sum_k \langle f, \Gamma_{j,k}(x) \rangle \Gamma_{j,k}(x) \vec{Y}} + \underbrace{\sum_{t=1}^j \sum_k \langle f, \Theta_{t,k}(x) \rangle \Theta_{t,k}(x) \vec{D}} \quad , \quad (3.5)$$

onde:

- $\Gamma(x)$ e $\Theta(x)$ são ortonormais e formam uma base para escrever $f(x)$, sendo a área sobre $\Gamma(x)$ unitária;
- $\vec{Y}$ e $\vec{D}$ são as projeções de $f(x)$ nos sub-espços V e W , respectivamente;
- V_i e W_i , os sub-espços *coarse morph* e *fine morph*, respectivamente, são ortogonais e $V_{i+1} = V_i \oplus W_i$;

3.1.3 A Morphlet de um sinal temporal discreto

Exatamente como na DWT, a *Morphlet* de um sinal temporal discreto $f_{[\cdot]}$ de suporte X é definida como sendo um outro sinal discreto de mesmo comprimento, sendo X uma potência de 2. A partir de $f_{[\cdot]}$ é possível obter uma *Morphlet* de nível no máximo j , sendo $j = \log_2(X)$, $j \in \mathbb{Z}$. O primeiro nível de decomposição de $f_{[\cdot]}$ gera dois outros sinais de suportes $\frac{X}{2}$ que são chamados *master morph signal* e *second-rated morph signal*. O primeiro resulta da convolução de $f_{[\cdot]}$ com $p_{[\cdot]}$ seguido de um *downsampling* por 2, enquanto que o segundo corresponde ao sinal obtido da convolução de $f_{[\cdot]}$ com $q_{[\cdot]}$, seguido do mesmo processo de *downsampling*, ambos submetidos ao processo de *wrap-around*. A concatenação do *master* com o *second-rated* corresponde à *Morphlet* de $f_{[\cdot]}$. A partir do *master*, a decomposição pode continuar recursivamente até o maior nível possível. É também admissível, assim como para qualquer DWT, a transformada *Morphlet-packet* de $f[\cdot]$, que pode ser calculada conforme explicado durante a revisão bibliográfica.

3.1.4 Cálculo da *Morphlet* e da *Morphlet* inversa

O cálculo da *Morphlet* e sua inversa pode ser realizado utilizando o algoritmo de Mallat (12), exatamente como descrito na seção 2.1.2 do Capítulo anterior. Para o cálculo, observa-se que o algoritmo requer somente o uso dos filtros $p_{[\cdot]}$ e $q_{[\cdot]}$.

3.2 O algoritmo de conversão de voz baseado na *Morphlet*

O algoritmo descrito a seguir permite utilizar uma *Morphlet*, especialmente construída para converter um certo sinal de entrada $f_{[\cdot]}$, pertencente ao *source speaker*, para o equivalente ao ditado pelo *target speaker*. Para todas as conversões, fica definido que os tamanhos de $f_{[\cdot]}$ e dos $s_{i[\cdot]}$ e $t_{i[\cdot]}$ é de 131072 amostras, com uma taxa de amostragem de 16000 amostras por segundo, no formato WAV. Esses dados caracterizam arquivos de vozes com cerca de 8 segundos de duração. Caso as gravações ultrapassem esse tamanho, o restante é descartado. Caso as durações sejam menores, será aplicado um processo de *zero-padding* (14). Além disso, todos os sinais são normalizados em amplitude antes do uso.

• INÍCIO

- **Passo P_1 :** criar uma *Morphlet*, isto é, um conjunto de filtros, considerando o conjunto de $s_{i[\cdot]}$ e os respectivos $t_{i[\cdot]}$ como sendo os exemplos de “treinamento” para que a *Morphlet* possa “aprender” a função de transferência entre o *source* e o *target speaker* ;
- **Passo P_2 :** calcular a *Morphlet* nível máximo, isto é, $\frac{\log(131072)}{\log(2)} = 17$, de $f_{[\cdot]}$;
- **Passo P_3 :** ortogonalizar o resultado da *Morphlet*($f_{[\cdot]}$). Isso pode ser feito usando a mesma relação existente para bancos de filtros QMF, que é utilizada para definir a relação entre os filtros $p_{[\cdot]}$ e $q_{[\cdot]}$ mencionados acima;

- **Passo P_4 :** inverter a transformada *Morphlet* anterior, já ortogonalizada. O resultado corresponde a voz convertida, podendo haver uma re-normalização, no sentido de amplificar o volume, para tornar a voz convertida audível;
 - **FIM.**
-

Os resultados da aplicação do algoritmo acima em sinais de voz encontram-se no próximo Capítulo.

Capítulo 4

Testes e Resultados

Neste capítulo encontram-se os testes realizados com o algoritmo de conversão de voz proposto no capítulo 3, variando-se o número de exemplos de treinamento utilizados para construir cada Morphlet, assim como os suportes dos filtros. Particularmente, os testes avaliam a qualidade das vozes convertidas, que é medida em termos de distâncias Euclidianas e também perceptualmente. As sentenças utilizadas são de conteúdo fonético rico, contendo trechos voiced e unvoiced, assegurando a experimentação para variados casos. Todas as vozes foram amostradas a uma taxa de 16000 amostras por segundo, 16-bits, sem compressão.

4.1 Testes de comparação dos resultados com distância Euclidiana

Na presente seção são apresentados os testes realizados e os resultados obtidos levando em conta a distância Euclidiana entre os sinais, original e convertido. A distância Euclidiana, d , entre dois vetores $x[\cdot]$ e $y[\cdot]$, ambos de comprimento c , é dada pela equação 4.1.

$$d = \sqrt{\sum_{i=0}^{c-1} (x_i - y_i)^2} \quad (4.1)$$

Os testes comparativos das vozes convertidas com suas versões originais encontram-se nos gráficos das figuras 4.1-A e 4.1-B e nas figuras 4.2-A e 4.2-B, onde SW é o suporte *Wavelet*, ou seja, o comprimento dos filtros da *Morphlet* e DI é a distância Euclidiana. O comprimento dos filtros *Morphlet* variou de 4 a 96. A descrição dos valores encontra-se nas tabelas 4.1 e 4.2.

Tabela 4.1- Distância Euclidiana entre a sentença *sal.wav* do diretório */timit/test/dr1/mdab0* e a *sal.wav* do diretório */timit/test/dr1/mjsw0*.

[M → M]:

SW	DI	SW	DI	SW	DI	SW	DI	SW	DI	SW	DI
-	-	4	1810	6	1810	8	1806	10	1806	12	1800
14	1800	16	1801	18	1768	20	1776	22	1750	24	1754
26	1758	28	1749	30	1730	32	1728	34	1727	36	1728
38	1722	40	1718	42	1711	44	1701	46	1699	48	1701
50	1690	52	1694	54	1690	56	1693	58	1660	60	1656
62	1643	64	1644	66	1632	68	1628	70	1620	72	1612
74	1611	76	1601	78	1588	80	1583	82	1512	84	1500
86	1489	88	1470	90	1468	92	1455	94	1444	96	1436

[F → F]:

SW	DI	SW	DI	SW	DI	SW	DI	SW	DI	SW	DI
-	-	4	1910	6	1909	8	1906	10	1886	12	1850
14	1840	16	1831	18	1828	20	1826	22	1801	24	1804
26	1757	28	1748	30	1722	32	1721	34	1721	36	1720
38	1722	40	1717	42	1711	44	1702	46	1696	48	1690
50	1690	52	1693	54	1690	56	1695	58	1664	60	1653
62	1647	64	1646	66	1633	68	1622	70	1618	72	1611
74	1610	76	1602	78	1583	80	1580	82	1547	84	1513
86	1499	88	1474	90	1473	92	1465	94	1433	96	1430

No gráfico da figura 4.1-A, a convergência foi analisada para a sentença *sal.wav* do diretório */timit/test/dr1/mdab0* da base TIMIT (26), do *Linguistic Data Consortium*, utilizada como *source speech*, para ser convertida na sentença *sal.wav* do diretório */timit/test/dr1/mjsw0*, da mesma base e de mesmo conteúdo textual, sendo que ambos os locutores são do sexo masculino. Seguindo o mesmo procedimento, o gráfico da figura 4.1-B contém os resultados obtidos quando a sentença *sal.wav* do diretório */timit/test/dr1/faks0* é convertida para a *sal.wav* do diretório

Tabela 4.2- Distância Euclidiana entre a sentença *sa1.wav* do diretório */timit/ test/ dr1/ mwbt0* e a *sa1.wav* do diretório */timit/ test/ dr1/ fjem0*.

[M $\rightarrow$ F]:

SW	DI	SW	DI	SW	DI	SW	DI	SW	DI	SW	DI
-	-	4	1602	6	1600	8	1593	10	1593	12	1590
14	1578	16	1574	18	1541	20	1535	22	1534	24	1521
26	1504	28	1506	30	1498	32	1487	34	1489	36	1488
38	1482	40	1484	42	1476	44	1468	46	1466	48	1458
50	1438	52	1428	54	1419	56	1416	58	1403	60	1402
62	1400	64	1401	66	1387	68	1384	70	1381	72	1373
74	1371	76	1269	78	1265	80	1261	82	1256	84	1252
86	1244	88	1199	90	1189	92	1182	94	1180	96	1176

[F $\rightarrow$ M]:

SW	DI	SW	DI	SW	DI	SW	DI	SW	DI	SW	DI
-	-	4	1701	6	1663	8	1632	10	1610	12	1609
14	1580	16	1573	18	1515	20	1513	22	1505	24	1500
26	1482	28	1462	30	1467	32	1439	34	1429	36	1408
38	1401	40	1383	42	1358	44	1331	46	1321	48	1315
50	1313	52	1310	54	1298	56	1271	58	1231	60	1229
62	1215	64	1199	66	1156	68	1155	70	1134	72	1123
74	1111	76	1102	78	1063	80	1088	82	1064	84	1032
86	1011	88	991	90	982	92	979	94	965	96	945

/timit/test/dr1/fadac1, sendo ambas de mesmo conteúdo textual, mas com ambos os locutores femininos. De acordo com os dois gráficos da figura 4.1, a maior distância Euclidiana foi obtida para o filtro *Morphlet* com suporte 4 e a menor distância Euclidiana foi obtida para o filtro *Morphlet* com suporte 96. Deve-se observar que, para um filtro *Morphlet* de suporte N , foi sempre utilizado 1 momento nulo e $\frac{N}{2} - 1$ exemplos de treinamento, isto é, pares de vetores $s_{[.]}$ e $t_{[.]}$. Todas as sentenças foram normalizadas em amplitude antes das medições das distâncias.

O gráfico da figura 4.2-A auxilia na descrição dos resultados obtidos quando a sentença *sa1.wav* do diretório */timit/test/dr1/mwbt0* é convertida para a *sa1.wav* do diretório */timit/test/dr1/fjem0*, sendo ambas de mesmo conteúdo textual, o *source speaker* masculino e o *target speaker* feminino. Finalmente, o gráfico da figura 4.2-B faz referência ao teste contrário, ou seja, quando a sentença *sa1.wav* do diretório */timit/test/dr1/fjem0* é convertida para a *sa1.wav* do diretório */timit/test/dr1/mwbt0*, caracterizando conversão de voz de um locutor de sexo feminino para o masculino. De acordo com os dois gráficos da figura 4.2, a maior distância Euclidiana foi ob-

tida para o filtro *Morphlet* com suporte 4 e a menor distância Euclidiana foi obtida para o filtro *Morphlet* com suporte 96. Novamenet, deve-se observar que, para um filtro *Morphlet* de suporte N , foi sempre utilizado 1 momento nulo e $\frac{N}{2} - 1$ exemplos de treinamento, isto é, pares de vetores $s_{[.]}$ e $t_{[.]}$. Todas as sentenças foram normalizadas em amplitude antes das medições das distâncias.

4.2 Testes de comparação dos resultados via critério perceptual

Os testes comparativos das vozes convertidas com suas versões pronunciadas pelos locutores originais, levando-se em conta a avaliação perceptual entre os sinais, original e convertido, encontram-se nos gráficos das figuras 4.3 e 4.4 onde SW é o suporte *Wavelet*, ou seja, o comprimento dos filtros da *Morphlet* e (M,D) correspondem a média e ao desvio-padrão, respectivamente. O comprimento dos filtros *Morphlet* variou de 4 a 96. Para essa comparação, 50 ouvintes voluntários, 25 do sexo masculino e 25 do sexo feminino, de idades entre 10 e 70 anos, foram questionados. Eles avaliaram cada uma das conversões numa escala de 1 até 5, onde 1 significa que a voz original não se parece em nada com a voz convertida, e 5 significa que ambas são idênticas, do ponto de vista perceptual.

O gráfico da figura 4.3-A serve para ilustrar os resultados dos testes realizados na conversão entre a sentença *sa1.wav* do diretório */timit/ test/ dr1/ mdab0* e a *sa1.wav* do diretório */timit/ test/ dr1/ mjsw0*, onde o *source speaker* e o *target speaker* são do sexo masculino. Seguindo o mesmo procedimento, o gráfico da figura 4.3-B ajuda na visualização dos resultados obtidos quando a sentença *sa1.wav* do diretório */timit/ test/ dr1/ faks0* é convertida na sentença *sa1.wav* do diretório */timit/ test/ dr1/ fadac1*, respectivamente, onde o *source speaker* e o *target speaker* são do sexo feminino.

De acordo com os dois gráficos da figura 4.3, a menor média foi obtida para o filtro *Morphlet* com suporte 4 e as maiores médias foram obtidas para os filtros *Morphlet* com suporte entre 44 e 96 com exceção de alguns filtros com suporte dentro desse intervalo. As tabelas 4.3 e 4.4 mostram os resultados detalhados. Mais uma vez, deve-se observar que, para um filtro de suporte N , foi sempre utilizado 1 momento nulo e $\frac{N}{2} - 1$ exemplos de treinamento, isto é, pares de vetores $s_{[-]}$ e $t_{[-]}$. Todas as sentenças foram normalizadas em amplitude antes das medições das distâncias.

Tabela 4.3- Conversão entre a sentença *sal.wav* do diretório */timit/ test/ dr1/ mdab0* e a *sal.wav* do diretório */timit/ test/ dr1/ mjsw0*.

[M $\rightarrow$ M]:

SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D
-	-	4	3,1	6	3,1	8	3,1	10	4,1	12	4,1
14	4,1	16	4,1	18	4,1	20	4,1	22	4,1	24	4,1
26	4,1	28	4,1	30	4,1	32	4,1	34	4,1	36	4,1
38	4,1	40	4,1	42	4,1	44	4,1	46	4,1	48	5,1
50	5,1	52	5,1	54	5,1	56	5,1	58	5,1	60	5,1
62	5,1	64	5,1	66	5,1	68	5,1	70	5,1	72	5,1
74	5,1	76	5,1	78	4,1	80	4,1	82	4,1	84	4,1
86	5,1	88	5,1	90	4,1	92	5,1	94	5,1	96	5,1

[F $\rightarrow$ F]:

SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D
-	-	4	3,1	6	3,1	8	3,1	10	4,1	12	4,1
14	4,1	16	4,1	18	4,1	20	4,1	22	4,1	24	4,1
26	4,1	28	4,1	30	4,1	32	4,1	34	4,1	36	4,1
38	4,1	40	4,1	42	4,1	44	4,1	46	4,1	48	5,1
50	5,1	52	5,1	54	5,1	56	5,1	58	5,1	60	5,1
62	5,1	64	5,1	66	5,1	68	5,1	70	5,1	72	5,1
74	5,1	76	5,1	78	5,1	80	4,1	82	4,1	84	4,1
86	5,1	88	5,1	90	4,1	92	5,1	94	5,1	96	5,1

O gráfico da figura 4.4-A serve para ilustrar os resultados dos testes realizados na conversão entre a sentença *sal.wav* do diretório */timit/ test/ dr1/ mwbt0* e a *sal.wav* do diretório */timit/ test/ dr1/ fjem0*, onde o *source speaker* é do sexo masculino e o *target speaker* é do sexo feminino. Seguindo o mesmo procedimento, o gráfico da figura 4.4-B auxilia na descrição dos resultados obtidos quando a sentença *sal.wav* do diretório */timit/ test/ dr1/ fjem0* é convertida na sentença

Tabela 4.4- Conversão entre a sentença *sal.wav* do diretório */timit/ test/ dr1/ mwbt0* e a *sal.wav* do diretório */timit/ test/ dr1/ fjem0*.

[M $\rightarrow$ F]:

SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D
-	-	4	3,1	6	3,1	8	3,1	10	4,1	12	4,1
14	4,1	16	4,1	18	4,1	20	4,1	22	4,1	24	4,1
26	4,1	28	4,1	30	4,1	32	4,1	34	4,1	36	4,1
38	4,1	40	4,1	42	4,1	44	4,1	46	4,1	48	5,1
50	5,1	52	5,1	54	5,1	56	5,1	58	5,1	60	5,1
62	5,1	64	5,1	66	5,1	68	5,1	70	5,1	72	5,1
74	5,1	76	5,1	78	4,1	80	4,1	82	4,1	84	4,1
86	5,1	88	3,1	90	4,1	92	5,1	94	5,1	96	5,1

[F $\rightarrow$ M]:

SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D	SW	M,D
-	-	4	3,1	6	3,1	8	3,1	10	4,1	12	4,1
14	4,1	16	4,1	18	4,1	20	4,1	22	4,1	24	4,1
26	4,1	28	4,1	30	4,1	32	4,1	34	4,1	36	4,1
38	4,1	40	4,1	42	4,1	44	4,1	46	4,1	48	5,1
50	5,1	52	5,1	54	5,1	56	5,1	58	5,1	60	5,1
62	5,1	64	5,1	66	5,1	68	5,1	70	5,1	72	5,1
74	5,1	76	5,1	78	4,1	80	4,1	82	4,1	84	4,1
86	5,1	88	3,1	90	4,1	92	5,1	94	5,1	96	5,1

sal.wav do diretório */timit/ test/ dr1/ mwbt0*, respectivamente, onde o *source speaker* é do sexo feminino e o *target speaker* é do sexo masculino.

De acordo com os dois gráficos da figura 4.4, a menor média foi obtida para o filtro *Morphlet* com suporte 4 e as maiores médias foram obtidas para os filtros *Morphlet* com suporte entre 44 e 96 com exceção de alguns filtros com suporte dentro desse intervalo. Também neste experimento, para um filtro de suporte N , foi sempre utilizado 1 momento nulo e $\frac{N}{2} - 1$ exemplos de treinamento, isto é, pares de vetores $s_{[-]}$ e $t_{[-]}$. Todas as sentenças foram normalizadas em amplitude antes das medições das distâncias.

Figura 4.1- acima: (A): masculino → masculino ; abaixo: (B): feminino → feminino

Figura 4.2- acima: (A): masculino → feminino ; abaixo: (B): feminino → masculino

Figura 4.3- acima: (A): masculino → masculino ; abaixo: (B): feminino → feminino

Figura 4.4- acima: (A): masculino → feminino ; abaixo: (B): feminino → masculino

Capítulo 5

Conclusões e Comentários

Este trabalho apresentou a criação de uma nova família de transformadas *wavelet*, chamadas de *Morphlets*, com aplicações para conversão digital de voz. Tanto a DWT criada quanto a técnica utilizada são novas na literatura e, claramente, forneceram uma contribuição original ao estado-da-arte na área.

Durante os testes, foi possível identificar que suportes maiores de filtros resultavam em funções de transferências muito mais funcionais para a finalidade desejada. Particularmente, filtros com suportes baixos, próximos de 4, implicavam no uso de um único exemplo de treinamento, isto é, um único par *source - target*, o que trazia resultados extremamente ruins, principalmente do ponto de vista perceptual. Já com suportes mais altos, por volta de 40 ou mais, os resultados perceptuais, principalmente, melhoravam consideravelmente. Nesse caso, conversões masculino-feminino, feminino-masculino, masculino-masculino e feminino-feminino mostraram-se possíveis e apresentaram resultados motivadores.

Este trabalho também introduziu na literatura o conceito de “treinamento de *wavelets*”, isto é, *wavelets* inteligentes. O uso desse termo da-se por analogia às redes neurais artificiais, que devem ser treinadas antes de realizar uma determinada classificação, por exemplo. Além disso,

não havia nenhuma menção na literatura sobre a possibilidade de realizar conversões, como a conversão de voz, sem o apoio de tais redes. Esse fato, portanto, também é novo.

O sistema descrito neste trabalho foi implementado em linguagem C/C++ sob ambiente Linux, porém, implementações para funcionamento em tempo-real, com base em um *Digital Signal Processor* (DSP) ou *Field Programmable Gate Array* (FPGA), também estão sendo planejadas.

Referências

- 1 VIEIRA, L.S. **Conversão de voz baseada na transformada wavelet**. 2007. 113p. Dissertação (Mestrado). Instituto de Física de São Carlos da Universidade de São Paulo, São Carlos, 2007.
- 2 WU, C. H.; HSIA, C. C.; LIU, T.H.; WANG, J.F. Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. **IEEE Trans. on Audio, Speech and Language Processing**, v.14, n.4, p.1109 - 1116, July 2006.
- 3 MOUCHTARIS, A.; VAN DER SPIEGEL, J.; MUELLER, P. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, (ICASSP '04), 17-21 May 2004, Quebec, Canada. **Proceedings..** .Quebec, Canada: IEEE, 2004. v.1., p. I 1-4.
- 4 GUIDO, R.C. **Spikelet**: uma nova transformada *wavelet* aplicada ao reconhecimento digital de padrões, em tempo-real, de *spikes* e *overlaps* em sinais neurofisiológicos do campo visual da mosca. 2003. 132p. Tese (Doutorado) Instituto de Física de São Carlos da Universidade de São Paulo, São Carlos, 2003.
- 5 GUIDO, R.C.; PEREIRA, J. C. Speechlet: uma nova transformada wavelet aplicada ao reconhecimento computacional de patologias em sinais da voz. In: III CISCi 2004 Conferência Iberoamericana de Sistemas, Cibernética e Informática, 2004, Orlando - FL - USA. v. 2. p. 191-195.
- 6 MALLAT, S. G. **A wavelet tour of signal processing**. San Diego: Academic Press, 1999.
- 7 STRANG, G.; NGUYEN, T. **Wavelets and filter banks**. Wellesley: Wellesley-Cambridge Academic Press, 1997.
- 8 ADDISON, P. S. **The illustrated wavelet transform handbook**: introductory theory and applications in science, engineering, medicine and finance. Edinburg-UK: Institute of Physics Publishing, 2002.

- 9 WILLIAMS, J. R.; AMARATUNGA, K. Introduction to wavelets in engineering, **International Journal for Numerical Methods in Engineering**, v. 37, n. , p. 2365-2388, 1994.
- 10 AKAY, M. **Time frequency and wavelets in biomedical signal processing**. New York: IEEE Press in Biomedical Engineering, 1998.
- 11 LIPSCHUTZ, S. **Álgebra linear**. 3.ed. São Paulo: Makron Books, 1994.
- 12 WALKER, J.S. **A primer on wavelets and their scientific applications**. Whashington: Chapman and Hall/CRC, 1999.
- 13 JENSEN, A.; COUR-HARBO, A. **Ripples in mathematics: The discrete wavelet transform**. New York: Springer-Verlag, 2000.
- 14 OPPENHEIN, A.V.; SCHAFER, R.W. **Discrete time signal processing**. 2. ed. New York: Prentice Hall, 1999.
- 15 GUIDO, R.C.; BARBON JR, S.; VIEIRA, L.S.; et al. Introduction to the discrete shapelet transform and a new paradigm: joint time-frequency-shape analysis. In: INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (ISCAS '08), 17-21 Jan 2008, Seatle, USA. **Proceedings..** .Seatle, USA: IEEE, 2008. v.1,, p. I 2893-2896.
- 16 GUIDO, R.C.; SLAETS, J.F.W.; KOBERLE, R.; ALMEIDA, L.O.B.; PEREIRA, J.C. A new technique to construct a *wavelet* transform matching a specified signal with applications to digital, real-time, spike and overlap pattern recognition. **Digital Signal Processing, Elsevier**, v. 16, n.1, p. 24-44, May 2006.
- 17 BARBON JR, S. **Dynamic time warping baseado na transformada wavelet**. 2007. 113p. Dissertação (Mestrado). Instituto de Física de São Carlos da Universidade de São Paulo, São Carlos, 2007.
- 18 ARSLAN, L.; TALKIN, D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In: INTERNATIONAL EUROPEAN SPEECH, (EUROSPEECH '97), 17-21 May 1997, Rhodes, Germany. **Proceedings..** .Rodes, Germany: IEEE, 1997. v.1,, p. I 1347-1350.
- 19 STYLIANOU, Y.; CAPPE, O.; MOULINES, E. Statistical methods for voice quality transformation. In: INTERNATIONAL EUROPEAN SPEECH, (EUROSPEECH '95), 17-21 May 1995, Madri, Spain. **Proceedings..** .Madri, Spain: IEEE, 1995. v.1,, p. 447-450.
- 20 CHOU, W.; JUANG, B.H. **Pattern recognition in speech and language processing**. Boca Raton: CRC Press, 2003.
- 21 KAIN, A.; MACON, M.W. Spectral voice conversion for text-to-speech synthesis. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, (ICASSP '98), 17-21 May 1998, Seatle, USA. **Proceedings..** .Seatle, USA: IEEE, 1998. v.1,, p. I 285-288.

- 22 ORPHANIDOU, C.; MOROZ, I. M.; ROBERTS, S. J. Voice morphing using the generative topographic mapping. In: INTERNATIONAL CONFERENCE ON COMPUTER, COMMUNICATION AND CONTROL TECHNOLOGIES, 17-21 May 2003, Orlando, USA. **Proceedings..** .Orlando, USA: IEEE, 2003. v.1,, p. I 222-225.
- 23 ORPHANIDOU, C. et al. Multiscale voice morphing using radial basis function analysis In: INTERNATIONAL CONFERENCE ON ALGORITHMS FOR APPROXIMATION LOCATION, 5,,: Chester, ENGLAND. **Proceedings ..** Chester, ENGLAND, Jul 17-21, 2005.p. 61-69 DOI: 10.1007/978-3-540-46551-5
- 24 DENG, L.; O'SHAUGHNESSY, O. **Speech processing:** a dynamic and optimization-oriented approach. New York: Marcel Dekker, 2003.
- 25 CLAUDIO, D. M.; MARINS, J. M. **Cálculo numérico computacional:** teoria e prática. 3. ed. São Paulo: Atlas, 2000.
- 26 TIMIT do Linguistic data consortium database. Disponível em: <http://www.ldc.upenn.edu/>. Acesso em Novembro/2011.

Apêndice I - Publicações durante o Doutorado

1. VIEIRA, L. S.; GUIDO, R. C.; J. C., CHEN, S. H. Introducing the Discrete Morphlet Transform and its Applications for Voice Conversion. In: 13th IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, (IEEE ISM 2011), 5-7 December 2011, Dana Point, USA. **Proceedings..** .Dana Point, USA: IEEE, 2011. v. 1,, p. 256-261.
2. BARBON JR, S.; GUIDO, R. C.; VIEIRA, L. S.; FONSECA, E.; SANCHEZ, F. L.; SCALASSARA, P. R.; MACIEL, C. D.; PEREIRA, J. C.; CHEN, S. H. Wavelet-based dynamic time warping. **Journal of Computational and Applied Mathematics, Elsevier**, v. 227, n.2, p. 281-287, May 2009.
3. SANCHEZ, F. L.; BARBON JR, S.; VIEIRA, L. S.; GUIDO, R. C.; FONSECA, E.; SCALASSARA, P. R.; MACIEL, C. D.; PEREIRA, J. C.; CHEN, S. H. Wavelet-based cepstrum calculation. **Journal of Computational and Applied Mathematics, Elsevier**, v. 227, n. 2, p. 288-293, May 2009.
4. SCALASSARA, P. R.; MACIEL, C. D.; GUIDO, R. C.; PEREIRA, J. C.; FONSECA, E. MONTAGNOLI, A.N.; BARBON, S.Jr.; VIEIRA, L.S.; SANCHEZ, F.L. Autoregressive Decomposition and Pole Tracking Applied to Vocal Folds Nodule Identification. **Pattern Recognition Letters, Elsevier**, v. 28, n.11, p.1360-1367, May 2007.
5. VIEIRA, L.S.; GUIDO, R. C.; BARBON, Jr, S.; SANCHEZ, F.L.; DUCATTI, J.A.; MOTA, T.B.; SANTOS, B.L.S.; HOU, T.H.; GOMES, M.H. (em Português) Identificação de Patologias por sinais da fala humana usando a transformada discreta de Fourier. **Boletim Técnico da Faculdade de Tecnologia de São Paulo (FATEC), São Paulo - SP**, v. 22, p. 53-56, Junho 2007.
6. GUIDO, R. C.; VIEIRA L. S.; BARBON JR, S.; SANCHEZ, F. L.; MACIEL, C. D.; FONSECA, E.; PEREIRA, J. C. A neural-wavelet architecture for voice conversion. **Neurocomputing, Elsevier**, v. 71, n.1-3, pp.174-180, September 2007.
7. GUIDO, R.C.; BARBON, S.J.; VIEIRA, L.S.; SANCHEZ, F.L.; MACIEL, C.D.; SCALASSARA, P.R.; PEREIRA, J.C.; PUIA, V.M. Spoken Document Summarization based on Dynamic Time Warping and Wavelets. **International Journal of Semantic Computing (IJSC), World Scientific**, v. 1, p. 347-357, 2007.

8. BARBON, S.Jr; GUIDO, R.C.; VIEIRA, L.S. A Neural-network approach for speech features classification based on paraconsistent logic. In: 11th IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (IEEE ISM 2009), 4-6 December, San Diego, USA. **Proceedings..** .San Diego, USA: IEEE, 2009. v.1., p. 567-570.
9. FANTINATO, P. C.; GUIDO, R. C.; CHEN, S. H.; SANTOS, B. L. S.; VIEIRA, L. S.; BARBON JR, S.; RODRIGUES, L. C.; SANCHEZ, F. L.; ESCOLA, J. P. L.; SOUZA, L. M.; MACIEL, C. D.; SCALASSARA, P. R.; PEREIRA, J. C. A Fractal-based Approach for Speech Segmentation. In: 10th IEEE IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (IEEE ISM 2008), 3-5 December, Berkeley, USA. **Proceedings..** .Berkeley, USA: IEEE, 2008. v.1., p. 551-555.
10. GUIDO, R.C.; BARBON JR, S.; VIEIRA, L.S.; et al. Introduction to the Discrete Shapelet Transform and a New Paradigm: joint time-frequency-shape analysis. In: IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS (IEEE ISCAS 2008), 18-21 May, Seattle, USA. **Proceedings..** .Seattle, USA: IEEE, 2008. v.1., p. 2893-2896.
11. BARBON JR, S.; GUIDO, R. C.; CHEN, S. H.; VIEIRA, L. S.; SANCHEZ, F. L. Improved Dynamic Time Warping Based on the Discrete Wavelet Transform. In: 9th IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (IEEE ISM 2009), 10-12 December, Taichung, Taiwan. **Proceedings..** .Taichung, Taiwan: IEEE, 2007. v.1., p. 256-261.
12. GUIDO, R. C.; VIEIRA, L. S.; BARBON, S.J.; SANCHEZ, F. L.; GUILHERME, M. B. A.; SEGIO, K. I. C.; SCARPA, T. L.; FONSECA, E.; PEREIRA, J. C.; MONTEIRO, M. A Fractal and wavelet-based approach for audio coding. In: 8th IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (IEEE ISM 2006), 11-13 December, San Diego, USA. **Proceedings..** .San Diego, USA: IEEE, 2006. v.1., p.253-256.
13. GUIDO, R. C.; MACIEL, C. D.; MONTEIRO, M.; FONSECA, E.; PANCHAPAGESAN, S.; PEREIRA, J. C.; VIEIRA, L.; BARBON, S.Jr.; SANCHEZ, F.L.; GUILHERME, M.B.A.; SÉRGIO, K.I.C.; SCARPA, T.L.; FANTINATO, P.C.; MOURA, J.R. A study on the best wavelet for audio compression. In: 40th IEEE ASILOMAR INTERNATIONAL CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS (IEEE ASILOMAR 2006), 15-17 July, Pacific-Grove, USA. **Proceedings..** .Pacific-Grove, USA: IEEE, 2006. v.1., p. 2115-2118.
14. GUIDO, R. C.; PEREIRA, J. C.; FONSECA, E.; SANCHEZ, F. L.; VIEIRA, L. S.; GUILHERME, M. B. A.; BARBON JR, S. Support Vector Machines and Wavelets for Voice Disorder Sorting. In 38th IEEE SOUTHEASTERN SYMPOSIUM ON SYSTEM THEORY (IEEE SSST 2006), 13-15 May, Cookeville, USA. **Proceedings..** .Cookeville, USA: IEEE, 2006. v.1., p. 434-438.
15. BARBON, L.S.; RODRIGUES, L.C.; VIEIRA, L.S.; GUIDO, R.C. Uso de Janelas Triangulares no Auxílio à Percepção Sonora por Portadores de Necessidades Especiais Au-

- ditivas. In: XXXIII CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL 2010 (CNMAC 2010), 20-23 Setembro, Águas de Lindóia, Brasil. **Proceedings..** Águas de Lindóia, Brasil: SBMAC, 2010. CD-ROM.
16. GUIDO, R.C.; JUNIOR, S.B.; VIEIRA, L.S.; FANTINATO, P.; SANCHEZ, F.L.; ESCOLA, J.P.L.; BOLOGNINI, E.; SOUZA, L.; ZULATO, P.; RIBEIRO, J.; LACERDA, M. Speech Compression based on the Discrete Shapelet Transform. In: XXXI CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL 2008 (CNMAC 2008), 8-12 Setembro, Belem, Brasil. **Proceedings..** Belem, Brasil: SBMAC, 2008. CD-ROM.
 17. BARBON, S.J.; VIEIRA, L.S.; SANCHEZ, F.L.; GUIDO, R.C. *Wavelets e Aplicações em Dynamic Time Warping*. In: 6th CONGRESS OF THE AUDIO ENGINEERING SOCIETY (AES 2007), 29-30 May, São Paulo, Brazil. **Proceedings..** São Paulo, Brazil: AES, 2007. CD-ROM.
 18. VIEIRA, L.S.; SANCHEZ, F.L.; BARBON, S.J.; GUIDO, R.C. On the Use of Wavelets for Audio Conversion. In: 6th CONGRESS OF THE AUDIO ENGINEERING SOCIETY (AES 2007), 29-30 May, São Paulo, Brazil. **Proceedings..** São Paulo, Brazil: AES, 2007. CD-ROM.
 19. SANCHEZ, F.L.; GUIDO, R.C.; VIEIRA, L.S.; BARBON, S.J. Wavelet-based Audio Watermarking. In: 6th CONGRESS OF THE AUDIO ENGINEERING SOCIETY (AES 2007), 29-30 May, São Paulo, Brazil. **Proceedings..** São Paulo, Brazil: AES, 2007. CD-ROM.
 20. GUIDO, R. C.; BARBON JR, S.; VIEIRA, L. S.; RODRIGUES, L. C.; FANTINATO, P. C.; ESCOLA, J. P. L.; SOUZA, L. M.; ALMEIDA, L. O. B.; SLAETS, J. F. W. Autenticação biométrica por comando de voz com implementação em *hardware* dedicado. In: II WORKSHOP NO ÂMBITO DO PROJETO - AMBIENTES PARA CO-PROJETO *HARDWARE/SOFTWARE* EM PLATAFORMAS FPGAs COM APLICAÇÕES EM ROBÓTICA MÓVEL (WAP 2008), 3 Abril, São Carlos, Brasil. **Proceedings..** São Carlos, Brasil: ICMC, 2008. CD-ROM.
 21. SANCHEZ, F.L.; GUIDO, R.C.; VIEIRA, L.S.; BARBON Jr, S. Inserção e Extração de Marca d'água em arquivos digitais de audio. In: XXX CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL (CNMAC 2007), 3-6 Setembro, Florianópolis, Brasil. **Proceedings..** Florianópolis, Brasil: SBMAC, 2007. CD-ROM.
 22. VIEIRA, L.S., GUIDO, R.C., BARBON Jr, S., SANCHEZ, F.L., SERGIO, K.I.C., GUILHERME, M.B.A., SOUZA, L.M., FANTINATO, P.C., BASSI, R.D.S., AMADIO, B.C.G., FONSECA, E.S. On the use of wavelets for voice morphing. In: XXX CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL (CNMAC 2007), 3-6 Setembro, Florianópolis, Brasil. **Proceedings..** Florianópolis, Brasil: SBMAC, 2007. CD-ROM.

23. VIEIRA, L.S., GUIDO, R.C., HOLAK, L.A., ZULATO, P., SANTOS B.L.S., MOTA, T.B., HOU, T.H., BARBERO, M.H.G., FEDERISSIS, J.F., PUIA, V.M., FERREIRA, A.A., VINHOTI, R.D. Reconhecimento de Vogais usando wavelets. In: XXX CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL (CNMAC 2007), 3-6 Setembro, Florianópolis, Brasil. **Proceedings..** .Florianópolis, Brasil: SBMAC, 2007. CD-ROM.
24. VIEIRA, L.S., GUIDO, R.C., LACERDA, M.A., RIBEIRO, J., SANTOS B.L.S., MOTA, T.B., HOU, T.H., BARBERO, M.H.G., FEDERISSIS, J.F., PUIA, V.M., FERREIRA, A.A., VINHOTI, R.D. Verificação de Locutor Usando as Transformadas Wavelet e Fourier. In: XXX CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL (CNMAC 2007), 3-6 Setembro, Florianópolis, Brasil. **Proceedings..** .Florianópolis, Brasil: SBMAC, 2007. CD-ROM.
25. VIEIRA, L. S.; GUILHERME, M. B. A.; GUIDO, R. C.; SERGIO, K. I. C.; ESCARPA, T. L. On the use of wavelets for image compression. In: XXIX CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL (CNMAC 2006), 18-21 Setembro, Campinas, Brasil. **Proceedings..** .Campinas, Brasil: SBMAC, 2006. CD-ROM.